Journal of Industrial and Systems Engineering
Vol. 15, No. 3, pp. 83-102
Summer (July) 2023

JISE

Prediction of Risk Factors in Cyber Harassment Using Big Data Analytics on social media

Mohammad Hossein Darvish Motevalli[1], Fariba Salahi[2*], Ladan Riazi[3], Adel Pourghader Chobar[4]

[1]*Department of Industrial Management, West Tehran Branch, Islamic Azad University, Tehran, Iran*

[2]*Department of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran*

[3]*Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran*

[4]*Department of Industrial Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran*

Abstract

Taking into account the fact that social media are today known as a platform to express and relieve the emotions, stress, and concerns that adolescents face in their daily lives, the ground has been provided for self-seekers. This is to the extent that these websites have been raised as a place for serious social problems and that vulnerable people, especially adolescents, are harassed on the Internet, commit suicide, or become bullies who harm others. Many new research papers are published every day in which various artificial intelligence (AI) techniques are applied to various tasks and applications related to sentiment analysis.

Keywords: Sentiment analysis, Support vector machine, Algorithm, Grey wolf optimization

1. Introduction

Given the considerable advancement of Internet technology, social media sites have become very popular and play a significant role in the evolution of human life. The extensive growth in popularity of online social networks in the last decade has unfortunately led to an unprecedented increase in the threat of cyberbullying (Muneer and Fati, 2020; Rahmaty et al. 2022). Thus, presenting research that provides proper insight into the analysis of Internet incidents and disturbances has become a major need in using social networks. Cyber harassment is an aggressive behavior carried out using electronic tools repeatedly and over time by a group or an individual against a victim who cannot easily defend themselves (Smith,

* Corresponding author
ISSN: 1735-8272,

2012). Cyber harassment is classified into seven types: flaming, cyber harassment, humiliation, impersonation, division, deprivation, and cyberstalking (Holland et al., 2009; Aliahmadi et al. 2013).

Moreover, the volume of data transferred in social networks has increased exponentially and the value of data is recognized as an economic asset. Accordingly, efforts are being made to actively use "big data" in various areas. Analysis of data mining decision tree, which uses big social data, is a useful tool for analysis and mutual effect (Hinduja and Patchin, 2010; Nozari etal. 2016). Along with all the positive effects the Internet and social networks have for people in society, their negative effects such as cyber harassment, internet addiction, preoccupation with games, etc. are indubitable. Considering the importance of these negative effects and their destructive effects on different people—in particular people of more sensitive ages—predicting risk factors in cyber harassment on social media has become an important issue. In this regard, in this research, an attempt is made to propose a model for predicting risk factors in cyber harassment on social media using the big data available in social networks and also using the methods and algorithms of data mining science.

2. Literature Review

Finding hidden patterns in messages exchanged on the Internet and social media can be widely used to predict risk factors in Internet harassment. In this way, the use of data mining methods in this field has attracted the attention of researchers. Therefore, collecting and reading comments on the web has become a necessity.

Hematpoor et al. (2017), conducted a study on the identification and sociological explanation of violence against women in cyberspace. The findings of the research showed that women's social participation and family attachment reduces violence against women in cyberspace and violence increases with the increase of online addiction, and conformity in cyber groups is effective in violence against women in the case of online addiction. The results of the regression analysis also indicated that the social participation variable is more than other variables affecting violence against women.

In a study by Muneer and Fathi, (2020), a solution to detect cyberbullying without involving the victims is proposed. To do this, a dataset containing 37,373 tweets from Twitter was used. In this research, seven classes of machine learning were used, namely, logistic regression (LR), light gradient boosting machine (LGBM), stochastic gradient descent (SGD), random forest (RF), AdaBoost (ADB), naïve Bayes (NB), and support vector machine (SVM). Each of these algorithms was evaluated using different performance metrics to determine the rate of dataset recognition. Experimental results show the superiority of logistic regression in this research. In a study by Herodotou et al., (2020), the first practical and real-time framework for aggression detection through pattern discovery using machine learning is presented. An aggression detection framework for social media is developed, in which the whole processing line from pre-processing and feature extraction to training uses the data flow model. The machine learning flow model is increasingly updated and can investigate new aggressive behaviors. Meanwhile, this framework is remarkably scalable because it can process the entire Twitter messages with only three machines. At the end of the results of this research, it was shown that the proposed flow framework can be easily used to detect other types of offensive behavior such as sarcasm, racism, and sexism with minimal effort.

In a study by Chatzakou et al., (2019), a powerful method for identifying bullies and aggressors from normal Twitter users was introduced, which identifies people considering text, user, and network-based attributes. In this research, using different advanced machine learning algorithms, the accounts were classified with more than 90% accuracy. Finally, the current status of Twitter user accounts that were identified as high-risk individuals with the proposed method was discussed and the performance of

potential mechanisms that can be used by Twitter to suspend users in the future was investigated. In another study, there is a new approach for timely and accurate detection of cyberbullying on Instagram. In this system, a sequential hypothesis testing formulation is used that seeks to drastically reduce the number of features used in classifying each comment while maintaining high classification accuracy. It raises an alert only after a certain number of detections have been made. Extensive experiments on a real-world Instagram dataset with ~ 4M users and ~ 10M comments demonstrate the effectiveness, scalability, and timeliness of our approach and its benefits over existing methods (Chelmis and Zois, 2019). Also, in a study, random forest was used to classify offensive comments in three cases of cyber aggression, including racism, sexual violence, and violence against women. The results of this research show that the accurate classification of comments related to cyber aggression can significantly reduce this phenomenon (Gutiérrez-Esparza et al. 2019). Nelson et al. (2018), researched cyber harassment: bullying and trolling. Half of people bullied online do not know who is behind it. Cyber harassment involves directing derogatory or offensive comments to targeted individuals. It can take various forms such as cyberbullying, cyberstalking, trolling, or spreading hate, for example.

In a study on incidents of cyberbullying in image- and video-based social networks, respectively, materials were collected and then a system solution that uses the insights obtained to improve the efficiency and effectiveness of cyberbullying detection was presented. The results of this research showed that the use of text and video features together greatly improves the classification performance of cyberbullying detection (Rafiq et al. 2018). Moon et al. (2011), conducted a study on the role of technology in youth harassment and the results showed that 60% of young Internet users experienced harassment in the past year and 11% experienced cyberbullying.

In a study by Singhal and Bansal, (2013), a survey on the current scenario of cyberbullying and the various methods available to detect and prevent cyberbullying is presented. In this research, a social networking website was designed and implemented, through which bullying was simulated and users were prevented from getting bullied. In addition, in this research, a case study was conducted and the performance of the system was evaluated. In a study, Mitchell et al. (2007), investigated the prevalence and frequency of Internet harassment. Of the young people who commit Internet harassment, 6% committed occasional acts of harassment, and 17% committed limited acts of Internet harassment in this country. Rule-breaking problems were reported three times more frequently. In general, behavioral and psychosocial problems increased in prevalence as the intensity of harassing behavior increased.

3. Methodology

The cross-industry standard process for data mining (CRISP-DM) methodology is one of the most popular and common methods for conducting data mining projects, which is used in this research. The main reason for choosing the CRISP-DM methodology over other methods, in addition to its comprehensibility for all users, is that this methodology is more widely used than other methods based on research conducted by the Data Mining Research Association.

The CRISP-DM methodology has six steps (Saadollahi et al. 2015). These steps include system recognition, data recognition, data preparation, modeling, evaluation, and development. Figure 1 shows how to perform these steps. This methodology is a loop, iterative, and interactive model. This means that some of its steps may be repeated and implemented several times to achieve the expected results in modeling. The CRISP-DM methodology is one of the most powerful methodologies in the field of data mining project implementation (Shafique and Qaiser, 2014).

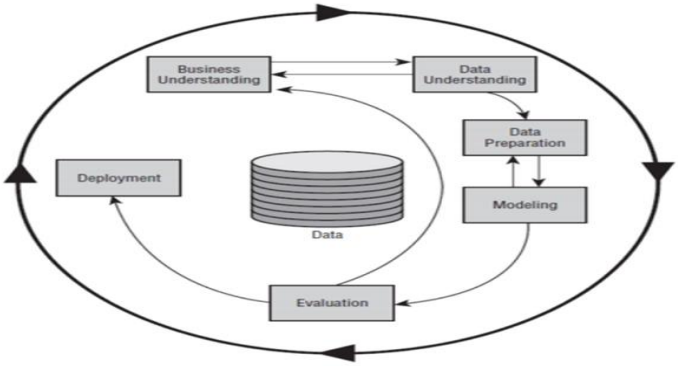

Figure 1. The CRISP-DM methodology phases (Shafique and Qaiser, 2014)

Stage 1 in CRIPS-DM: System Recognition

This stage includes collecting the requirements and information required about the system and interviewing senior managers and experts. In fact, at this stage, the system is recognized and the objectives and key success factors of the system are determined and revised (Ameri et al. 2013; Bathaee et al. 2023). This research aims to use the big data available on social media and also to use the methods and algorithms of data mining science to predict risk factors in cyber harassment on social media. Figure 2 shows the various steps performed in the first stage.

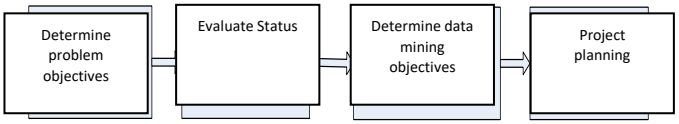

Figure 2. Stage 1 in CRIPS-DM, system recognition

Stage 2 in CRIPS-DM: Data recognition

The second stage of the CRISP-DM method is related to the concept of data and is dedicated to data recognition. This stage includes primary and main data collection, data description, data exploration, data quality research, and data collection (Saadollahi et al. 2015). Figure 3 demonstrates the different steps of the second stage.

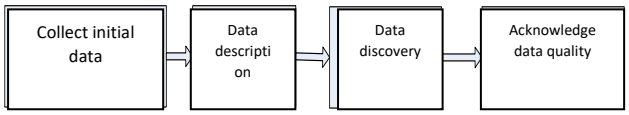

Figure 3. Stage 2 in CRIPS-DM, data recognition

Stage 3 in CRIPS-DM: Data preparation

Data preparation is one of the most important and often the most important tasks in data mining projects and includes data selection, data cleaning, data transformation, new data structuring, and data integration. As stated previously, today, due to the large amount of data and the connection with different information sources, information banks are exposed to the existence of contradictory, confusing, and missing data. Using preprocessing and data preparation techniques, it is possible to increase the quality of the data and consequently the quality of the output results. The preprocessing process considered in this research is shown in Figures 4 and 5.

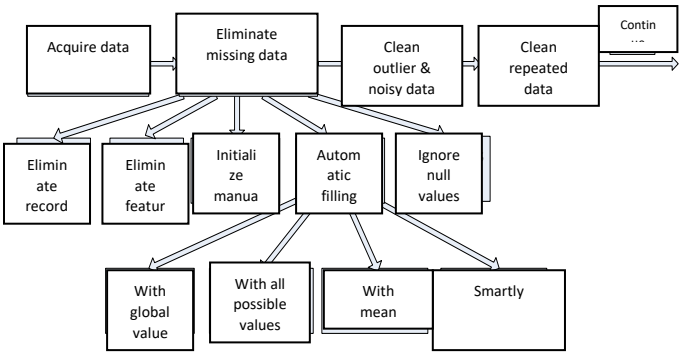

Figure 4. Preprocessing process

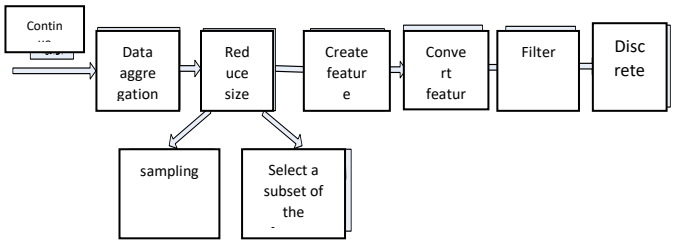

Figure 5. The rest of the preprocessing process

Stage 4 in CRIPS-DM: Modeling

At this stage, the desired model is found using different data mining techniques. The input of the modeling state is the output of the preparation stage; therefore, the backward movement can take place to achieve a quality model. The different steps of this stage are displayed in Figure 6.

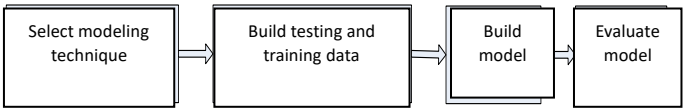

Figure 6. Stage 4 in CRIPS-DM: Modeling

There is no best among the algorithms and models, and the desired method should be selected according to the desired data and efficiency. There are many methods to learn the model, and usually each of them is powerful in a specific part, and to use one of them, necessary checks should be done to know how they work.

Stage 5 in CRIPS-DM: Evaluation criteria

The fifth stage in the CRISP-DM process focuses on evaluating the obtained models and deciding how to use the results. As shown in Figure 7, briefly in this section, the evaluation of the results, the review process, and the determination of the next steps are discussed. The interpretation of the model depends on the algorithm, and the models can be evaluated to check whether it fulfills the objectives properly or not. After that, the parts that did not reach the objectives will sometimes be repeated, or sometimes the initial objectives and settings will be changed. In order to evaluate data mining models, various criteria have been presented, some of these criteria are discussed below. Before introducing the evaluation criteria, it is

necessary to introduce the concepts of confusion matrix, true positive, false positive, true negative, and false negative.

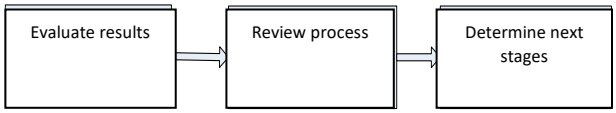

Figure 7. Stage 5 in CRIPS-DM: Evaluation criteria

In the field of artificial intelligence, the confusion matrix is a matrix in which the performance of the built model is displayed. Each column of the matrix shows the number of samples of classified values, if each line contains a real example (Jafari et al. 2012; Nozari and Ghahremani-Nahr et al. 2023). Figure 8 demonstrates the actual and predicted labels for the four outcomes of the confusion matrix (Saito and Rehmsmeier, 2015).

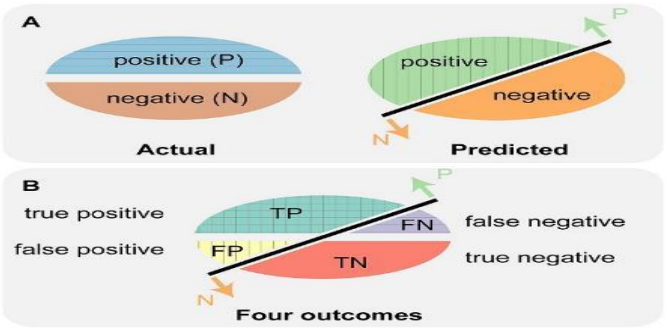

Figure 8. Actual and predicted labels for the four outcomes of the confusion matrix (Saito and Rehmsmeier, 2015)

The figure on the left in part A shows the two main labels: positive items (blue color) and negative items (red color). The figure on the right of part A shows the predicted cases: predicted positive items (light green), predicted negative cases (orange). The dashed line drawn in part A represents a classifier dividing the data into positive predicted items under the heading P and negative predicted under the heading N. In part B, the actual and predicted labels are shown in one figure. The result of this combination is the generation of the following four groups:

True positive (green color) indicates the items that are really positive and the classification method, correctly, identifies as positive.

False negative (purple color) indicates the items that are truly positive and the classification method could not identify them as positive.

False positive (yellow color) indicates the items that are not positive and the classification method has wrongly identified as positive.

True negative (red part) indicates the items that are not really positive and the classification method has correctly identified them as negative. Table 1 shows the primary evaluation criteria that can be extracted from the confusion matrix.

Table 1. Primary evaluation criteria that can be extracted from the confusion matrix (Saito and Rehmsmeier, 2015)

Measure	Formula
ACC	(TP + TN) / (TP + TN + FN + FP)
ERR	(FP + FN) / (TP + TN + FN + FP)
SN, TPR, REC	TP / (TP + FN)
SP	TN / (TN + FP)
FPR	FP / (TN + FP)
PREC, PPV	TP / (TP + FP)
MCC	$(TP * TN—FP * FN) / ((TP + FP)(TP + FN)(TN + FP)(TN + FN))^{1/2}$
$F_{0.5}$	1.5 * PREC * REC / (0.25 * PREC + REC)
F_1	2 * PREC * REC / (PREC + REC)
F_2	5 * PREC * REC / (4 * PREC + REC)

The concepts of some of these criteria are explained below:

The accuracy (ACC) criterion: This criterion is known as the accuracy of a model, which for a classification method is equal to the identified positive items divided by the total number of samples.

The sensitivity (SN) or true positive rate (TPR) criterion or recall: This criterion is known as recall rate, which for a classification method is equal to true positive divided by the sum of true positive and false negative.

The positive predictive value (PPV) criterion or precision: This criterion is known as recall rate, which for a classification method is equal to true positive divided by the sum of true positive and false positive.

Step 6 in CRIPS-DM: Development

As shown in Figure 9, this stage is related to how to use Stein's stage model to describe program expansion, program maintenance, final report generation, revision, and project publication.

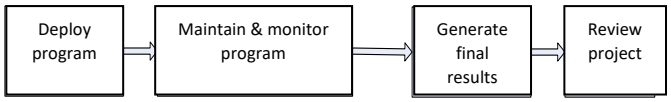

Figure 9. Stage 6 in CRIPS-DM: Development

Building a model is not the end of a project, and the goal of data mining projects is to discover knowledge and use the discovered knowledge in the future. The discovered knowledge should be organized and made available to others in a useful way. For this purpose, it is tried to determine the set of rules and important features extracted from the data set. This knowledge can be used to predict risk factors in cyber-harassment using big data analysis in social media.

3.1. Evaluation process in the proposed system

Preprocessing process

Before applying data mining algorithms on the data, preprocessing operations are performed on them. For example, one of the things that should be done is the removal of isolated letters, this part is related to the

removal of single letters because single letters do not help to extract patterns in the text mining process and only increase the length of the vector of extracted features.

Tokenization

The process of separating words based on a specific separator (usually a hyphen) is called tokenization. These words, numbers, and other specified characters are called tokens. In the tokenization stage, the search is summarized with the degree of importance. In this way, along with token generation, this process also evaluates the frequency value of all tokens in the input documents. The storage space is used to store the detected important tokens more appropriately. Tokenization is suggested in any preprocessing method in which the initial document vectors are dependent on the validity of tokens (Singh and Saini, 2014).

Stop Words

Stop words in text mining processes are words that are usually widely used in the language under investigation and are useless in the term. Stop words are part of natural language. The purpose of eliminating stop words from a text is to sort the words of the text and remove less important words in the analysis. In text documents, prepositions, etc. are the biggest stop words (Hadi et al. 2017).

Determination of TF-IDF

TF-IDF is a numerical statistic that indicates the importance level of a word relative to a document in a set of documents. Its purpose is to show the importance of the word in the text. The value of TF-IDF is balanced by the number of documents in the collection that contain the word, and it increases in proportion to the number of word repetitions in the document. It means that if a word appears in many texts, it is probably a common word and is not so valuable in evaluating the text (Dastani et al. 2020). In general, the importance of a word in a collection of documents is determined by two indicators: one is the relative frequency of occurrence of that word in the document, which is called term frequency (TF), and the other is the number of documents that contain that document, which is called document frequency (DF). DF expresses the ratio of documents containing that word in all documents. If the frequency of occurrence of a word in all documents is less than the existing document, it means that that word distinguishes the existing document better than other documents. To calculate them, first, the frequency of word i in document j (Fij) is calculated and by normalizing it in the whole collection, the value of TF is obtained, i.e.:

$$TFij=Fij/max(Fij) \quad (1)$$

IDF is based on the fact that words that appear in many documents are less expressive of the overall topic. For this reason, to calculate it, first the number of documents that contain the word i (ni) and the total number of documents in the collection (N) are determined, then IDF is calculated as follows:

$$IDFi=log(N/ni) \quad (2)$$

Finally, the TF-IDF weighting method is calculated as follows:

TF-IDF=Fij*IDFi (3)

This evaluation shows the product of TF and IDF and expresses the importance of a word in the document, and based on that, the words in the documents can be ranked according to their importance.

N-gram extraction

In text mining processes, n-gram means a connected sequence of words or words. For example, 3-gram means sequences of three words or letters. Bi-gram means two words and it places the words in a sentence two by two next to each other, and the tri-gram model puts the words in a sentence three by three next to each other. N-gram simply displays all significant combinations of adjacent words or letters of length n found in a text. One of the important features that N-gram provides is that it captures the language structure statistically. The optimal length of n depends on the application and the context of its use. If the N-gram is too short, it may not record an important difference, and if it is too long, it only depicts general information (Taher et al. 2018).

Dividing data into training and testing to implement and check the efficiency of the proposed system, the data analyzed in this research are divided into training data and test data. In this division, the share of data used for training the system is 80% of the total data examined. The remaining 20% are used for system evaluation and are considered as test data.

Modeling

In this section, as shown in Figure 10, the full details of the general proposed model for implementation are explained. In the rest of this section, explanations are provided about the evaluation criteria used and the modeling methods examined.

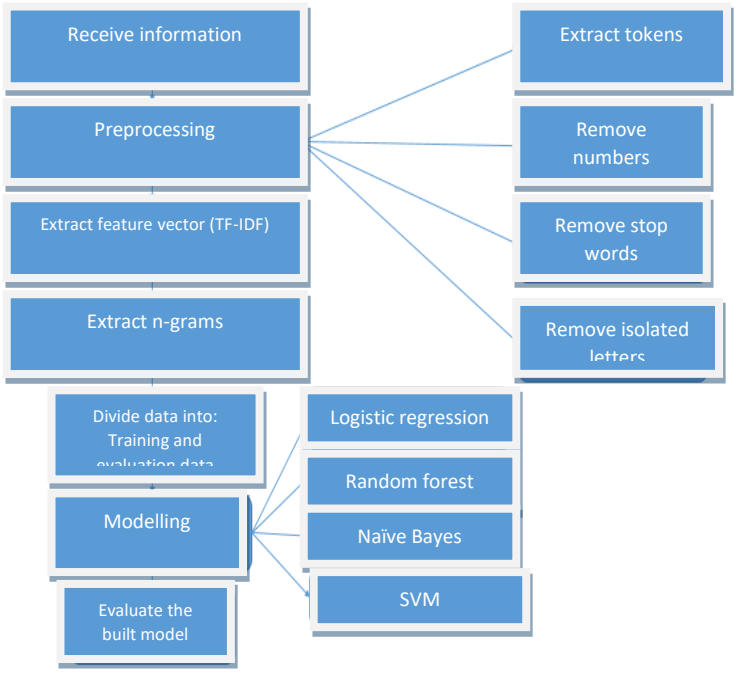

Figure 10. General proposed model for implementation

Evaluation criteria

In this section, various criteria used in data mining research will be discussed. Some of these criteria will be used to evaluate the implementation of this research. For simplicity, explanations about the concepts of true positive, false positive, true negative, and false negative are given first. Then, definitions will be given for the various evaluation criteria used in this research, including accuracy, precision, recall, and F-measure criteria.

False positive (FP) rate is defined as the fraction of negative samples that are detected as positive.

$$False\ positive\ rate = \frac{FP}{FP + TN} \tag{4}$$

False negative or (FN) rate refers to the fraction of positive samples that are detected as negative.

$$False\ negative\ rate = \frac{FN}{TP + FN} \tag{5}$$

The true positive (TP) rate is the fraction of positive samples that are correctly classified as positive.

$$True\ positive\ rate = \frac{TP}{TP + FN} \tag{6}$$

True negative or (TN) rate refers to the fraction of negative samples that are correctly classified as negative.

$$True\ negative\ rate = \frac{TN}{TN + FP} \tag{7}$$

The accuracy criterion is known as the accuracy of a model, which for a classification method is equal to the identified positive items divided by the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

Sensitivity (SN) or true positive rate (TPR) or recall is known as recall rate, which for a classification method is equal to true positive divided by the sum of true positive and false negative.

$$Recall\ = \frac{TP}{TP + FN} \tag{9}$$

Positive predictive value (PPV) or precision is known as recall rate, which for a classification method is equal to true positive divided by the sum of true positive and false positive.

$$\text{Precision}\ = \frac{TP}{TP + FP} \tag{10}$$

The F-measure is obtained from the combination of the precision and recall criteria and is used in items in which it is not possible to assign special importance to each of the two precision and recall criteria.

F-Measure is expressed in the form of Equation (11).

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

Logistic regression models used

The regression model is a statistical model in which the relationship between a phenomenon (dependent variable) and some of its factors (independent variables) is defined based on a series of observed values. Normal logistic regression describes the relationship between a two-state response variable (presence or absence of a variable) and a set of response variables (Xn,...,X2,X1). The response variables may be continuous or discrete and do not need to be distributed frequently. The logistic response function can be written as the following equation:

$$P(Y_i = X_i) = \frac{Exp(B_1 + B_2 + X_i + \cdots + B_n + X_i)}{1 + Exp(B_1 + B_2 + X_i + \cdots + B_n + X_i)} \quad (12)$$

Where p is the probability of occurrence, Yi is the dependent variable, Xi is the independent variable, B1 is the constant coefficient, B2 is the angle coefficient of the variables (y-intercept) and EXP is the exponential function (Nosrati et al., 2018). One of the advantages of using the logistic regression model, in addition to modeling the observations, is the possibility of predicting the probability of each individual belonging to each of the levels of the dependent variable, as well as the possibility of directly calculating the likelihood ratio using the coefficients of the model (Sedehi et al. 2010; Hosseinzadeh-Lotfi et al. 2016).

Random forest classifier

Random forest is actually a cumulative learning method for classification, regression, and other tasks. Random forests work by building a large number of decision trees during training, and its output is classification, which can mean creating classes in samples (classification) or predicting their types (regression). Random decision forest actually corrects the habit of over-fitting in decision trees.

Random forest is created by growing a large number of classification (decision) trees. To classify a new object from an input vector, the classification of the tree will be determined for that object and the class to which the object belongs in terms of the tree will be recorded. To calculate the permutation importance index, the random forest algorithm does not use all the observations of the sample to build a tree, but a random sample is selected by placing it in the size of n1 (usually equal to $^{2n}/_{3}$). The selected observations are called the learning sample (LS) and the rest of them are called the out-of-the-bag (OOB) sample. Trees are built with LS observations and OOB is used to measure tree impurity. In each tree, the impurity size is first calculated on the OOB observations. Then, the values of variable xi in OOB observations are randomly permuted and tree impurity size is calculated on the permuted values. The importance of variable xi in each tree is the difference between these two impurity sizes and the mean of these values is the permutation importance index. The motivation of this method is that if xi is an important variable, randomly permuting its values will lead to an increase in the impurity of the tree, while if it is not an influencing variable, there will be no change in the impurity (Noori et al. 2011; Mehrani et al. 2019).

Naive Bayes

The naive Bayes classification method is one of the Bayes classifications and is designed based on Bayes' law. In this method, the probability that the sample belongs to each category is calculated, and accordingly,

the record is assigned to the category with the highest probability. For this purpose, P(Y|X) is calculated, which, as we know, according to Bayes' rule, is obtained as follows:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)} \qquad (13)$$

where X is the set of values corresponding to the attributes of the test new record. $\prod_{i=1}^{d} P(X_i|Y)$, considering this point, it was found that in the naive Bayes classifier, we assume that the probabilities P(Xi|Y) are independent of each other. Considering that P(X) is a constant value, it is enough to choose class Y so that the form of the above fraction is maximized. It means that we try the classes one by one to find the class with the highest result in this expression. To calculate P(Xi = a|Y = b), we must obtain the result of dividing the number of samples whose class is b by the number of records in which the characteristic Xi has a value equal to a (Braha, 2013).

Support vector machine

Support vector machine (SVM) was the first algorithm proposed by Fisher for classifying and categorizing models in 1936, and its criterion was to optimize and reduce the error of training data classification. SVMs have the following properties:

Designing a classifier with maximum generalization, reaching the overall optimal point of the function, automatically determining the optimal structure and topology for the classifier, modeling nonlinear discriminant functions using nonlinear kernels, and the concept of inner product in Hilbert spaces.

SVM is an algorithm that finds a specific type of linear model maximizing the margin of the hyperplane. Maximizing the margin of the hyperplane leads to the maximization of the separation between the classes. The closest training points to the maximum margin of the hyperplane are called support vectors. Only these vectors (points) are used to specify the border between classes (Shin et al. 2005).

4. Findings

To implement the proposed solution, there is a need for a dataset including people's opinions. Therefore, online platforms have been used to prepare the required data. One of these platforms is the well-known Snapptrip website, which offers travel and transfer services in its subseries. In this way, the collection of comments in the Snapptrip website has been used as the investigated dataset, and in this research, the Python programming language and the artificial intelligence libraries proposed in this language have been used.

In this section, the results of the implementation of this research are presented. Table 2 shows the result of system implementation with 2000 features and an n-gram equal to 1. In the first part of this table, which is marked as "Simple data mining method", data mining methods including logistic regression algorithm, random forest, naive Bayes and SVM are implemented by default. In the second part, i.e., the part named "Data mining method with Spark", the mentioned methods have been implemented in a distributed manner. In this table, the implementation results were compared in terms of accuracy, precision, recall, F-measure, and execution time in each of the methods.

The results obtained in Table 2 indicate that the execution time of all the investigated algorithms is reduced if they are executed as Spark. Another result obtained from this table is that the SVM algorithm has a better overall performance in the evaluation criteria, while the random forest algorithm has provided the weakest results.

Table 3 shows the result of system implementation with 2000 features and an n-gram equal to 2. According to the results obtained from this table, when compared with the n-gram setting equal to 1 and the same number of features, the accuracy of the system and the execution time of the algorithm have not changed significantly in different modes.

Table 2. Results obtained from different methods with 2000 features and Ingram equal to 1 in simple mode and Spark

Kaggle	Tfidf	ngram = 1	numFeatures=2000		
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:58.19	82.15	81.69	55.69	66.23
Random Forest	0:00:59.72	78.57	70.95	53.88	61.25
Naive Bayes	0:00:56.01	79.94	81.05	47.19	59.65
SVM	0:01:02.62	81.93	78.86	58.04	66.87
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:04.17	82.25	81.96	55.23	66.79
Random Forest	0:00:04.58	78.56	70.32	54.78	62.07
Naive Bayes	0:00:11.75	80.03	81.66	46.99	59.22
SVM	0:00:49.65	81.84	78.56	58.06	66.18

Table 3. Results obtained from different methods with 2000 features and Ingram equal to 2 in simple mode and Spark

Kaggle	Tfidf	ngram = 2	numFeatures=2000		
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:58.35	81.76	81.52	54.24	65.14
Random Forest	0:00:59.74	78.40	70.25	54.24	61.22
Naive Bayes	0:00:57.16	79.88	80.24	47.73	59.86
SVM	0:01:51.64	74.84	72.56	40.06	51.62
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:04.44	81.17	81.84	54.65	65.22
Random Forest	0:00:05.09	78.11	70.03	54.12	61.18
Naive Bayes	0:00:12.82	79.71	80.27	47.14	59.81
SVM	0:00:49.65	74.84	72.52	40.01	51.60

Table 4 exhibits the result of system implementation with 2000 features and n-gram equal to 3. According to the results obtained from this table, when compared with the n-gram setting modes equal to 1 and 2 and the same number of features, there is no significant difference in the evaluation criteria and algorithm execution time in these three modes.

Table 4. Results obtained from different methods with 2000 features and n-gram equal to 3 in simple mode and Spark

Kaggle	Tfidf	ngram = 3		numFeatures=2000	
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:55.12	81.64	81.42	53.88	64.85
Random Forest	0:00:54.01	79.82	80.18	47.55	59.70
Naive Bayes	0:00:58.73	79.43	72.68	55.33	62.83
SVM	0:01:03.94	81.36	78.19	56.41	65.54
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:02.65	81.36	81.34	53.82	64.85
Random Forest	0:00:03.19	79.88	80.10	47.24	59.22
Naive Bayes	0:00:10.24	79.23	72.16	55.72	62.32
SVM	0:00:49.65	81.41	78.07	56.55	65.47

Table 5 demonstrates the result of system implementation with 1000 features and ngram equal to 1. According to the results obtained from this table, if the number of features is set to 1000, the running time of the algorithm will be relatively less. By comparing the results of this mode with

the mode with 2000 features and n-gram equal to 1, it can be seen that reducing the number of features will also reduce the accuracy of the algorithm.

Table 5. Results obtained from different methods with 1000 features and n-gram equal to 1 in simple mode and Spark

N_Gram Effect with numFeatures=1000					
Kaggle	Tfidf	ngram = 1	numFeatures=1000		
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:52.43	81.59	81.02	54.06	64.85
Random Forest	0:00:51.66	78.86	71.80	53.88	61.57
Naive Bayes	0:00:50.16	79.20	80.06	45.02	57.63
SVM	0:00:59.86	81.36	78.92	55.51	65.18
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:03.96	81.55	80.89	54.02	64.92
Random Forest	0:00:04.69	78.26	71.83	53.44	61.30
Naive Bayes	0:00:12.72	79.26	79.95	44.99	57.37
SVM	0:00:49.65	81.65	78.56	55.81	65.17

Table 6 shows the result of system implementation with 1000 features and n-gram equal to 2. According to the results obtained from this table and its comparison with the mode of 2000 features and n-gram equal to 2, it can be seen that in addition to the increase in accuracy, a noticeable decrease in the algorithm execution time is also observed.

Table 6. Results obtained from different methods with 1000 features of an n-gram equal to 2 in simple mode and Spark

N_Gram Effect with numFeatures=1000					
Kaggle	Tfidf	ngram = 2	numFeatures=1000		
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:51.28	81.36	80.82	53.34	64.27
Random Forest	0:00:53.92	78.69	71.81	52.98	60.97
Naive Bayes	0:00:49.77	79.14	80.19	44.66	57.37
SVM	0:00:58.33	81.42	78.68	56.05	65.46
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:02.88	81.83	80.85	53.22	64.42
Random Forest	0:00:03.60	78.52	71.92	52.78	60.49
Naive Bayes	0:00:09.15	79.25	80.05	44.75	57.31
SVM	0:00:32.09	81.40	78.63	56.15	56.65

Table 7 indicates the result of system implementation with 1000 features and n-gram equal to 3. According to the results obtained from this table, compared to the mode where the number of features is equal to 2000 and n-gram is considered 3, the total execution time of the algorithm is reduced, especially in the Spark part, and the accuracy increases relatively.

Table 7. Results obtained from different methods with 1000 features and n-gram equal to 3 in simple mode and SparkN_Gram Effect with numFeatures=1000

Kaggle	Tfidf	ngram = 3	numFeatures=1000		
Simple data mining method					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:53.69	81.30	80.60	53.34	64.20
Random Forest	0:00:54.63	79.54	74.18	53.52	62.18
Naive Bayes	0:00:51.24	79.26	80.51	44.84	57.60
SVM	0:00:59.23	81.30	78.71	55.51	65.11
Data mining method with Spark					
Method	Time	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0:00:02.98	81.21	80.37	53.22	64.05
Random Forest	0:00:03.62	79.90	74.02	53.47	62.20
Naive Bayes	0:00:09.46	79.72	80.41	44.54	57.22
SVM	0:00:31.68	81.48	78.93	55.51	65.01

5. Discussion and Conclusion

The destructive effects of cyber harassment have led to the development of many automated and data-driven methods with an emphasis on the accuracy of classification. However, the importance of timely detection and warning about cyber harassment immediately after detecting an offensive message is considered of great importance. Therefore, the importance of timely detection is essential to support victims as soon as possible. In this research, using data from online social media called Kaggle Parsed dataset, a method of detecting harassment in social media was investigated by examining logistic regression, random forest, naive Bayes, and support vector machine algorithms, and implementation. Python programming language was used to preprocess the data and then classify them. Furthermore, the suggested algorithms based on Spark were implemented and evaluated. And in a distributed form in the Spark format, it has achieved the proper accuracy and response time.

References

Aliahmadi, A., Jafari-Eskandari, M., Mozafari, M., & Nozari, H. (2013). Comparing artificial neural networks and regression methods for predicting crude oil exports. International Journal of Information, Business and Management, 5(2), 40-58.

Ameri, H., Alizadeh, S., & Barzegari, A. (2013). Knowledge extraction from the data of diabetic patients using the decision tree method. Journal of Health Management, (53)16.

Bathaee, M., Nozari, H., & Szmelter-Jarosz, A. (2023). Designing a new location-allocation and routing model with simultaneous pick-up and delivery in a closed-loop supply chain network under uncertainty. Logistics, 7(1), 3.

Braha, D. (2013). Data mining for design and manufacturing: methods and applications. Springer Science & Business Media.

Chatzakou, D., Leontiadis, I., Blackburn, J., Cristofaro, E. D., Stringhini, G., Vakali, A., & Kourtellis, N. (2019). Detecting cyberbullying and cyberaggression in social media. ACM Transactions on the Web (TWEB), 13(3), 1-51.

Dastani, M., Mousavi Chalek, A., Ziyai, S., & Delghanadi, F. (2020). Thematic analysis of published articles of medical librarianship and information in Iran using text mining techniques. Health Image, 1(1), 10-15.

Gutiérrez-Esparza, G. O., Vallejo-Allende, M., & Hernández-Torruco, J. (2019). Classification of cyber-aggression cases applying machine learning. Applied Sciences, 9(9), 1828.

Hadi, R. M., Hashem, S. H., & Maolood, A. T. (2017). An effective preprocessing step algorithm in text mining application. Engineering and Technology Journal, 35(2B), 232-241.

Hematpoor, B., Mazaheri, A. M., & Mohseni, R. A. (2017). Detection and sociological explanation of violence against women in cyber space. Women's Research Journal, 8(4), 131-105.

Herodotou, H., Chatzakou, D., & Kourtellis, N. (2020). A streaming machine learning framework for online aggression detection on Twitter. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 5056-5067). IEEE.

Hinduja, S., & Patchin, J. W. (2010). Bullying, cyberbullying, and suicide. Archives of Suicide Research, 14(3), 206-221. https://doi.org/10.1080/13811118.2010.494133

Holland, D., Ireland, J. L., & Muncer, S. (2009). Impulsivity, attribution and prison bullying: Bully-category and perpetrator–victim mutuality. International journal of law and psychiatry, 32(2), 84-91. https://doi.org/10.10/j.ijlp.2009.01.00

Jafari, E., & Samadian, M.-a. (2012). The use of data mining in investigating the behavior of delinquent drivers in big cities. Traffic Studies Quarterly, 9(17), 14-35.

Lotfi, F. H. Z., Najafi, S. E., & Nozari, H. (Eds.). (2016). Data envelopment analysis and effective performance assessment. IGI Global.

Mandela, N., & Kennedy, J. F. (2018). Online harassment: bullying, stalking and trolling. In Internet Literacy Handbook.

Mehrani, K., Mirshahvalad, A., & Abbasi, E. (2019). Comparison of the Accuracy of Black Hole Algorithms and Gravitational Research and the Hybrid Method in Portfolio Optimization. International Journal of Finance & Managerial Accounting, 4(14), 111-126.

Mehrani, K., Mirshahvalad, A., & Abbasi, E. (2019). Portfolio optimization using black hole meta heuristic algorithm. specialty journal of accounting and economics, 5(2-2019), 1-13.

Moon, B., Hwang, H. W., & McCluskey, J. D. (2011). Causes of school bullying: Empirical test of a general theory of crime, differential association theory, and general strain theory. Crime & Delinquency, 57(6), 849-877.

Muneer, A., & Fati, S. M. (2020). A comparative analysis of machine learning techniques for cyberbullying detection on twitter. Future Internet, 12(11), 187.

Noori, S., Nourijelyani, K., & Mohammad, K. (2011). Random Forests Analysis: A modern statistical method for screening in high-dimensional studies and its application in a population-based genetic. Journal of North Khorasan University of Medical Sciences, 3, 49-57.

Nozari, H., & Ghahremani-Nahr, J. (2023). A Comprehensive Strategic-Tactical Multi-Objective Sustainable Supply Chain Model with Human Resources Considerations. Supply Chain Analytics, 4, 100044.

Nozari, H., Najafi, S. E., Jafari-Eskandari, M., & Aliahmadi, A. (2016). Providing a model for virtual project management with an emphasis on IT projects. In Project Management: Concepts, Methodologies, Tools, and Applications (pp. 476-496). IGI Global.

Rafiq, R. I., Hosseinmardi, H., Lv, R., Han, Q., & Mishra, S. (2018). Scalable and timely detection of cyberbullying in online social networks. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (pp. 1738-1747). ACM.

Rahmaty, M., Daneshvar, A., Salahi, F., Ebrahimi, M., & Chobar, A. P. (2022). Customer churn modeling via the grey wolf optimizer and ensemble neural networks. Discrete Dynamics in Nature and Society, 2022(1), 9390768.

Saadollahi, A., & Najimeh, A. (2015). Data mining comparison of PSIRC methodology and RMEES methodology. In Proceedings of the Second International Research Conference in Science and Technology.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), e0118432.

Sedehi, M., Mehrabi, Y., Kazemnejad, A., & Hadaegh, F. (2010). Comparison of artificial neural network, logistic regression and discriminant analysis methods in prediction of metabolic syndrome. Iranian Journal of Endocrinology and Metabolism, 11(4), 638-646.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217-222.

Shin, S., Lee, S., & Kim, J. (2005). An application of support vector machines in bankruptcy prediction model. Expert Systems with Applications, 28(1), 127-135.

Singh, V., & Saini, B. (2014). An effective tokenization algorithm for information retrieval systems. Department of Computer Engineering, National Institute of Technology Kurukshetra, Haryana, India.

Singhal, P., & Bansal, A. (2013). Improved textual cyberbullying detection using data mining. International Journal of Information and Computation Technology, 3(6), 569-575.

Smith, P. K. (2012). Cyberbullying and cyber aggression. In Handbook of school violence and school safety (pp. 93-103). Routledge.

Taher, S. A., Akhter, K. A., & Hasan, K. A. (2018). N-gram based sentiment mining for Bangla text using support vector machine. In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-5). IEEE.

Yao, M., Chelmis, C., & Zois, D. S. (2019, May). Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In The World Wide Web Conference (pp. 3427-3433).

Ybarra, M. L., & Mitchell, K. J. (2007). Prevalence and frequency of Internet harassment instigation: Implications for adolescent health. Journal of Adolescent Health, 41(2), 189-195.