International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 11 (2019) pp. 219-226

Received: 15 Oct, 2019; Accepted: 18 Dec 2019; Publish: 21 Dec 2019

Toxic Comments Identification in Arabic Social Media

Osama Hosam

The collage of computer science and engineering in Yanbu,
Tiabah University, Saudi Arabia.
As with SRTA-City, IRI institute, Alexandria, Egypt.
mohandesosama@yahoo.com

***Abstract*: The usage of social media increases day by day. Both individuals and organizations use social media for different purposes. Problems increase in association with social media technologies. Toxic comments bots create a negative impression about people, companies and products. These kinds of toxic comment bots are created by the attackers. This research work is carried out to identify these toxic comments in Arabic social media. For that, Machine learning techniques are used. Mainly gradient boosting technique (XGBoost algorithm) has been utilized to effectively identify the comments created by the toxic comment bots. XGBoost can efficiently divide toxic comments into the following categories, toxic, severe toxic, obscene, threat, insult, and identity hate. The accuracy achieved by the proposed method is 99.54 %.**

***Keywords*:** *malware detection; machine learning; XGBoost , Adaboost , Classification, Clustering;*

I. Introduction

In the modern information era, people share their information on various platforms. Among them, social media sites and social networking sites play an important role. The use of social media sites and social networking sites is increased drastically [1]. Almost all people in the globe use social media as well as social networking sites. Companies use social media sites for advertising. Social Media platforms are a major source of personal data. It is an avenue for the attackers who are involved in the cyber-attacks etc. The most common type of malware attack on the social media website is comments based attacks. These attacks are mainly carried out to increase the traffic of porn and illegal websites. And, attackers damage the company or a famous person's name. For that, they develop toxic comments bots. It creates toxic commands frequently. Mainly the company's uses social media pages to promote the products and services. [2,3]

The specially created bots spoil the products name by using negative comments and offensive words. [4] These offensive words similar to the words of human when they are in higher angry and dissatisfaction. When the user reads these comments, they can easily get a negative impression on the product or services. In the same manner, it also creates problems for celebrities. Here the attackers create the bot to spread fake comments and pornographic pictures and videos of the celebrity in the comment section. These are termed as toxic comments.

Toxic comments are nothing, but the comments contain the offensive language or hurt someone or negative comments. Toxic comments may contain the threatening comments, obscene, identity-based hatred and insulting comments. These are considered as online harassment. Because of these actions, most of the peoples get confused and judging the peoples and products wrongly. Also, it creates a negative thought about the Arabic social media sites. In some cases, these activities also lead to psychological problems for celebrities. [5]

In Arabic social media company's perspective, these issues create a negative impact on the site. So, the people hesitate to use social media site. The company lost its valuable customers. In some cases, peoples also file a case against the company. There is no permanent solution to this problem. We only restrict these kinds of malicious activities on a social media site. For that companies and researchers continuously work on this area to develop the most viable solution to this problem. [6] There are many pieces of research are conducted on this topic by many peoples. [7] Different peoples founded different things and different solutions to these problems. [5] But there is no single solution that fits for all the cases. This proposed research intended to develop the model for classifying the toxic comments. Because identification of the toxic comments helps to identify the bot.

In the long run, scientists and researchers try to deploy machine learning and artificial intelligence technologies to find toxic comments bot from social networking and social media sites. Machine learning is one of the major applications of AI. [8] Adoption of machine learning for technologies for finding the toxic comments bot from the internet provides the ability to automatically learn and improve from the experience. This process doesn't require any explicit programs etc. In machine learning techniques initially, the example situations are provided to the model. Machine learning model observes the patterns and various data present in the given example. [9] Using the collected information the system identifies the factors which impact on the decision. The major aim of these kinds of system developments is to make self-learning computers. In general machine learning technologies are classified into four different technologies. And they have supervised machine learning algorithms, unsupervised machine learning algorithms, semi-supervised machine learning algorithms and reinforcement machine learning algorithms. [10] All these techniques have their own advantages and disadvantages. But one thing is common in all

these techniques. And it is data preparation for the analysis. Data preparation or data cleansing or data organizing process consumes more time and energy than the data analysis process in most of the cases.
The proposed systems have some limitations. The major limitation is error diagnosis and correction. Finding the error and correcting the error in the machine learning model is not an easy task. There is a huge number of complications are involved in this process. Also boosting techniques requires some time to learn. It is not possible to make the decision immediately. When compared to other machine learning algorithms boosting algorithm make decisions based on the historical data it has. See Figure 1., so it needs time. Immediate implementation is not possible. Also, it can solve only specific kind of problems. This is the most common limitation present in the machine learning approach. Also, the proposed research aims to reduce the above-discussed limitations. But it is the secondary aim of the paper. The primary aim of the paper is to develop the classification model.
Gradient boosting algorithm is one of the most powerful and widely used machine learning algorithm. Because of its powerful features, most of the researchers use boosting algorithms for their research. Boosting has the potential to make modifications as per the requirement. Because of their higher flexibility, boosting algorithms are mostly preferred for different practical problems. Especially data processing. boosting plays a vital role in data preprocessing [11].
Mainly boosting algorithms are used for classification, regression and ranking etc. They have the potential to build a custom tree. In general boosting algorithm means an algorithm which has the potential to make the changes in the training data. In the boosting technique, some values are assigned to the dataset. These values are termed as dataset score. This value helps to find the difficulties involved in the classification process. Figure 1 shows the classification technique using boosting algorithm. Boosting trains weak classifiers and come up with new more accurate classifier.

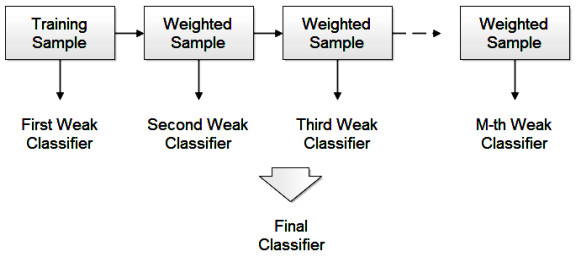

Figure 1. The boosting algorithm trains on a collection of weak classifiers which result in the more accurate classifier.

In this paper, XGBoosting will be adopted for the following reasons. First, it has higher scalability. So, XGBoosting suites for analyzing dataset irrespective of size. Second, it can solve the problem of missing values efficiently. In machine learning, missing values are the major problem. It influences the accuracy of the decision made. The proposed algorithm is good enough to deal with the missing values. Third, XGBoosting is very robust. Its ability to deal with irrelevant input data is too efficient.
However, boosting algorithms in general have some limitations. For extracting the linear combination from the dataset, it is not the most successful method. It won't provide higher accurate results on that. And it has a lower predictive power [13]. XGBoosting is more powerful and can handle the traditional boosting algorithm limitations.

The main objective of the paper is to identify the toxic comments bot present on Arabic social media sites from the negative comments it makes in social media websites. For that, the toxic comment classification model needs to be created. The developed model must be capable of reading the comments present in the dataset. Then it needs to split the comments into different classifications. And finally, the model must be identifying the toxic comment bots which creates negative comments.
The paper is organized as follow, section 1 contains the introduction. It brings a clear idea about the background of the proposed approach. Also, it describes the requirements of the project. Section 2 research methodology and research techniques and tools used. And the findings and discussions are provided in section 3 of the paper. Finally, the conclusion contains a brief overview of the methodology and findings of the research. Also, it contains the future works needed and recommendations etc.

II. Background

This section brings the basic idea about the methodologies in literature and practical difficulties faced by the different researchers.

A. Conventional Bots Identification methodologies

There are two types of methods for identifying toxic comment bots. They are behavior-based and signature-based. There is static and dynamic analysis. Static analysis is always permanent. They have no executed files. The dynamic method is performed when the files are executed. Static analysis can read the source program of the bot. They have the characteristics of the file and they can search into those characteristics [12]. There are different techniques in static analysis. (1) Examining metadata: Data about data (meta-data) in the file can give the main details in file format inspection. (2) Program output: Program Execution indicates the identification of the output data and files. It can gather details of the bot. (3) Fingerprinting: Fingerprinting contains cryptanalysis by hash calculations. When the file contains bot, anti-virus scanners can find it. For this, we use AV scanning. (4) Disassembly: Disassembly technique is used in reversing the computer program to assembly language and gathering the logic of software and thus that logic for the bot can be analyzed.

B. XGBoosting algorithms

XGBoosting is on type of sparsity-aware algorithms [13] it is utilized for sparse information. Its major purpose is tree learning. In such algorithms, there are cache access patterns, in addition to compression of data and fragmentation to create a tree boosting method.
In their paper [6], the author explains XGBoost algorithm for text classification. It is an algorithm that is used for machine learning. Tree boosting is one of the most significant machine learning technique. The algorithm XGBoost uses fewer resources than other approaches. It can resolve real-world scale problems by using fewer resources
XGBoosting is an open-source machine learning package. XGBoosting has scalability in nature. This algorithm is depicted in the field of engineering as a decision tree method.

It supports the following interfaces such as CLI, C++, JAVA & JVM languages. It enables parallel computing which provides rapid learning that allows fast model exploration. The algorithm can convert a large amount of data into fewer resources. XGBoosting can manage all sparsity patterns in a unified direction. Tree models are simple and accurate models. They have higher prediction accuracy.

Online comments have many positive and negative effects on public fields. Sometimes, they are beneficial but some other times they will create problems. Researchers spend most of their time for gathering, cleaning and forming the data.

Toxic comment classification is one of the modern fields and various studies have been seen to classify toxic comments. Logistic regression, Naïve Bays with Support Vector Machine (NBSVM), XGBoost and FastText algorithm with Bidirectional LSTM (FasText-BiLSTM) are the four commonly used classification algorithms [14]. Logistic regression is utilized by many researchers for Twitter comments [15] Logistic regression is not affected when erasing whitespaces. There are almost 35 ways for data transformation. These transformations will lead to higher accuracy. Twitter data has less character count while comment data has more character count. Therefore, toxic data is not balanced. Wang and Manning used a classification algorithm that depends on Naïve Bayes (NB) and Support Vector Machine (SVM). They obtained good results as NB introduced better performance with short text and SVM introduced higher accuracy with relatively long text [16]. FastText is an open-source package [17]. It has good memory capacity and faster than other algorithms. BiLSTM is the improved version of (Long Short-Term Memory) LSTM.

XGBoost is scalable and accurate boosting algorithm. XGBoost is used by many researcher's text classification. Its implementation is comparatively new. ML competitions used XGBoost in their winning approaches [6].

C. Deep Learning Algorithms

In their paper [18], the author gets information from the website communication page which includes many types of toxicity in online comments. There are many problems faced by online groups. Tormentor and online persecution are two of the major problems. There are various kinds of data extension methods. This method is used to recover the imbalanced problems of the data. The solution to this issue is a group of 3 models. They are CNN, LSTM, and GRU.

The classification methods can be categorized into two methods. In the first method, we should define the input is toxic or not. Secondly, we should detect the toxic types seen in the content. From this, the author says that the assembled method performs better in other algorithms. This study shows that CNN has high precision on different levels. CNN is a multi-label program because the input has multilevel toxicity. Nowadays, social communications are very popular in our world. It's very important to share relevant information and social communications. This platform will help us to express our concepts and views.

There are many psychological problems occurred in our society due to these social harassments. Fear, abuse, indecent words, and identity hate are some of the problems. In this paper, individuals can tackle issues by using data extension methods. The multi-label classification system can identify different kinds of toxicity.

A granulated classification of various kinds as well as aims of online hate as well as other learning models. These models are used to find and categorize the intolerable comments in the set of information and also testing with machine learning. It must consist of Decision trees, Random forest, Adaboost to produce a grouping design that must find and classifies toxic comments. Although various techniques had been utilized to decrease hateful comments. It contains comments, non-privacy as well as compulsory registration. There are various disadvantages to toxic comments. There are online firestorms. The bad comment had the ability to fright away high-class discusser prepared to provide positive comments in the discussion. It improves the polarity of a given set and echoes slot effects. Hateful comments contain overlying objectives and kinds of language, stimulating for multilabel categorization. YouTube experts do not give details about the state during comment range, 34.9% are from the US. Classification had initiated with the given rules. Finding themes, take the significance of comment, etc. Initially, the granulated classification of bad online commentaries. Then there is a multilabel design that must be used to group bad commentaries. At last, detect the given tasks for finding online bad speaking. There are complexities in explanation as well as the strength of views of bad speaking. It alters opinions between different persons, variation in language, restrictions of mechanization, etc. [19].

III. Methodology Overview

The main idea is to find negative information spreading social media bots by finding toxic comments. First, the toxic comments will be identified. Then they are classified into verities. And then the bot which created the comments are identified. For performing the classification process, gradient boosting technique will be employed. Mostly majority of peoples used other techniques like bagging and random forest models etc. [20] for their approaches. But their techniques are not capable to improve themselves. That's the main reason behind the selection of the boosting technique in our approach. Boosting techniques are capable of learning from previous failures. So, the accuracy of the model increases subsequently. By using the gradient boosting technique, the dataset is classified into different classes. After that, the regression process is done by the XG boosting technique. The developed system must read the comments presented in the CSV file. And then it processes the data and classifies the data into different types based on the toxicity. It is expected to split the comments into three types. (High toxicity comments, medium toxicity comments and low toxicity comments). Figure 2 shows the overview of the processes that will be carried out in the proposed approach.

A. Data Gathering

The data is mainly collected from social media websites such as, Facebook, Twitter, Instagram, WhatsApp, etc. The crawler collected both positive and negative posts and tweets. The positive entries are those entries having bad comments, the negative entries are bad-comments free. The most difficult part, here is the labeling process. The labeling process are done basically depending on the existence of some words, these words are called "poison keywords list". Examples of poison keywords list are "corrupted", "tyranny", and "filthy" in Arabic language. We reached in some cases to labeling

more than 90% of the entire dataset. The remaining entries are labelled manually. The labelling divides the positive entries into the following categories, toxic, severe toxic, obscene, threat, insult, and identity hate. We put logic 1 in the corresponding entry if it belongs to one of the previous six categories, otherwise we put logic 0.

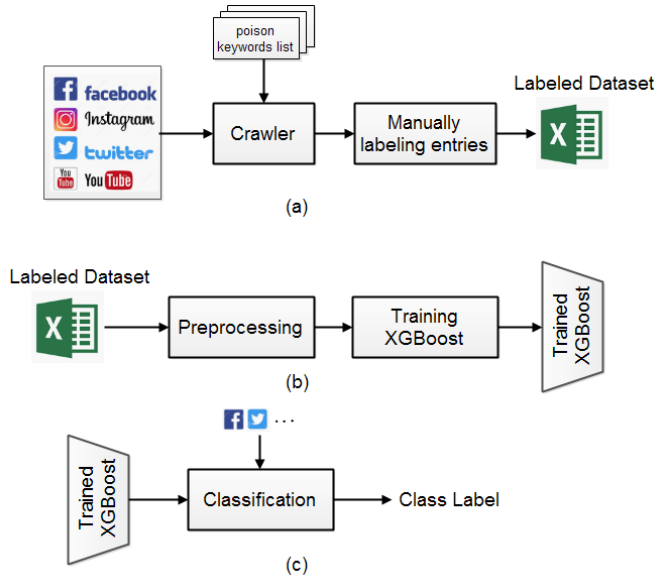

Figure 2. Methodological overview of the developed model. The figure explains the different stages involved in the process. (a) Data gathering (b) Model training (c) Classification

B. Model Training

This section contains information about the techniques used in the training phase. As already stated, the boosting algorithm is used in the training and testing phases [21], so it will be explained.

1) Preprocessing

The collected-labeled dataset is preprocessed before passing it to training. The data is cleaned from noisy entries, such as the entries collected by the crawler and have no relationship to the toxic comment detection study. Some missing values will force us to remove the entire entry from the collected dataset. Duplicate records are deleted by a measure such as distance. If the distance value is near zero, the entry is removed.

2) Gradient Boosting Algorithm

Gradient boosting machines (GBM) are a group of machine learning estimators [22]. The final output is the serial decision done by different estimators. Boosting is different from bagging. In bagging, different classifiers make decisions and the final decision is chosen to be the average output of the different classifiers. However, in boosting classifier decision is fine-tuned (boosted) by new classifier, the strong classifier remains, and the weak classifier is neglected. Simply saying, Bagging is parallel process while boosting is serial one.

Gradient boosting is utilized to produce precise models. This method must be empirically verified itself to be more effective for a huge array of categorization as well as regression problems. It is a new version of ensemble technique [23]. The probability is joined from various predictors. The goal of this technique is to train a set of decision trees. The training of individual decision tree is called Apriori. This method is termed as boosting. The aim of this method is to decrease the loss of the classifier model by increasing a weak learner at one time.

Boosting is an algorithm that must be globally utilized in the area of machine learning. This algorithm sets a weak classifier to weighted type. At every repetition, the data is reweighted. The miscategorized information points gets high weights. In boosting scheme, different weak learners must be joined. The strong learner gets great precision. The main factors utilized to calculate the precision of the algorithm are bias as well as variance. The best algorithm gives great bias as well as less variance

The objective of the learning process is to define a loss function and trying to minimize the loss. Let's say we are using mean squared error (MSE) as our loss function:

$$\text{Loss} = \text{MSE} = \sum(y_i - y_i^p)^2 \quad (1)$$

Where y_i is the ith target value, y_i^p is the ith prediction. We want the predicted values to minimize the loss or get minimum MSE. The gradient descent is used to reduce the loss in each iteration. The predictions are updated according to the learning rate, and hence we can find those values with minimum loss.

$$y_i^p = y_i^p + \frac{a \times \delta \times \sum(y_i - y_i^p)^2}{\delta \times y_i^p} \quad (2)$$

Which becomes,

$$y_i^p = y_i^p - \alpha \times 2 \times \sum(y_i - y_i^p) \quad (3)$$

Where, α is the learning rate and $\sum(y_i - y_i^p)$ is the sum of residuals.

We are updating the learning predictions to the sum of the residuals are minimized. The residuals are minimized only if the output is near the expected values.

In GBMs the learning process keeps the latest and strong models. It is greatly associated with the negative gradient of the loss function. The loss functions used must be arbitrary. If the error function is the squared-error loss - as stated before, then the learning process must outcomes into error fit. [24].

In function approximation, the learning is managed, and it leaves a great limitation on the investigator. [25] The information had to be given with an adequate group of target labels. The main changes among boosting techniques as well as conservative machine learning methods are that improvements must be done in the function space.

The extension for gradient boosting is the extreme gradient boosting. XGBoost supports greedy method. The ensemble is constructed serially. K-trees are utilized to group illustrations into classes.

Use of XGBoosting algorithms provides the below-listed advantages.

- *Regularization*: Regularization controls the overfitting. There are hyperparameters which are added to equation (1) to control overfitting.
- *Parallel Processing*: XGBoost uses parallel processing. GBM is slower than XGBoost. To perform the model, it utilizes more CPU cores.
- *Handling Missing Values*: XGBoost can manage missing values. It uses both hand split to meet missing value at a point. It ensures the same when functioning on testing data.

The speed, as well as performance, are high in XGBoost. In comparison with gradient boosting, XGBoost had high performance because of parallel processing. It is also scalable.

It is utilized in various applications. It helps outdoor memory. It is utilized for classification, regression as well as ranking. It handles overfitting. It also provides good performance outcomes on various set of data. The loss function is defined as the changes among real as well as the predicted value. It also affects the precision. It is convex and provides two kinds of errors. They are a positive component error as well as negative component error. Negative component error decrements the precision. Outliers mean the error that must be physically produced in the set of data. It always decrements the performance. The robust loss function is defined as the conditional probability of a class label.

C. *Classification*

Data are collected from social media for testing purposes. Part of the dataset can also be used for verification. The collected data are passed to the trained XGBoost model. The XGBoost model is multi-classifier model, it classifies data into six categories. Namely, toxic, severe toxic, obscene, threat, insult, and identity hate. The trained model is tested on a variety of datasets collected randomly from Arabic social media.

IV. Results and Discussion

Equalize For implementation, the python programming language will be used. Python is one of the most versatile programming languages for machine learning. Python brings many features to carry out big data analysis, machine learning model developments. [26] Python has an extensive collection of built-in libraries. These libraries allow users to perform different processes like machine learning, image recognition, data mining and artificial intelligence etc. In this section, the dataset is first explored followed by the training phase and the classification phase.

A validation test set is used instead of training set to evaluate model accuracy. The test set is selected using cross-validation. K-fold is used with k=10 as typical and popular value. The dataset is divided into k groups each group is different from the other (mutual exclusive). Each subset or group is approximately of equal size. In each iteration, a collection of groups is selected as training set and the remaining as test set. The selection of group maybe stratified to add regularity to the randomness in subset selection.

The classifier accuracy is measured by accuracy and error rate. Using the definition in Table 1 the accuracy and error rate can be defined as

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (4)$$

And the error rate is defined as (1 – accuracy) or

$$Error\ rate = \frac{FP+FN}{TP+FN+FP+TN} \quad (5)$$

The precision also called (exactness) is what percentage of the number of entries that has been classified as positive are actual positive values.

$$precision = \frac{TP}{TP+FP} \quad (6)$$

The recall also called (completeness). Recall concentrates on positive samples only, then measure the percentage of the positive samples that classified as positive by the classifier. The optimal value of recall is 1.

$$recall = \frac{TP}{TP+FN} \quad (7)$$

There is an inverse relationship between precision and recall. There is another measure called F-score which measure the harmonic mean of the precision and recall. It is mathematically defined as

$$F\_score = \frac{2\times precision\times recall}{precision+recall} \quad (8)$$

Table 1: Actual class versus predicted class's cases

Actual class/ predicted class	**Insult**	**Not-Insult**
Insult	True Positive (TP)	False Negative (FN)
Not-Insult	False Positive (FP)	True Negative (TN)

A. *The Dataset*

The dataset is mainly collected in a spreadsheet. The spreadsheet contains 8 columns, the first column is the comment id, the second column is the comment text. Columns from the second to the 8th are comment classes. If the comment belongs to the class, logic 1 is assigned to that entry and the remain classes are assigned logic 0. Figure 3 shows sample of the toxic and clean comments. Comment with id ends with 777bf is clear text and the comment id ends with c4d57 is identity hate comment. The Arabic social media is versatile and contains both clean and toxic comments. We have successfully collected 159529 records. The labeling process was very hard as sometimes we are forced to make it manually.

id	COMMENT_TEXT
0000997932d777bf	تفسير
000103f0d9cfb60f	شكر. (نقاش) 21:51 ، 11 يناير D'2016
000113f07ec002fd	يا رجل ، أنا حقا لا أحاول تعديل الحرب.
0001b41b1c6bb37e	"
0001d958c54c6e35	أكثر
00025465d4725e87	لا يمكنني تقديم أي اقتراحات حقيقية
006cf8c9f4cc4d57	آسف ، الرابط الذي أعطيته لي ميت. وأنا
006d11791d76b9f3	فقط اعترف أن الفيتناميين كلهم حفنة من الناس المتملقين الحمقى. جميع شعوب شرق آسيا الذين تحدثت معهم يعتقدون أن فيتنام جزء لا يتجزأ من آسيا ، وعلى العكس أعتقد أنها دوله متخلفة وقذرة وتتحدث بلغة مزعجة.

Figure 3. Sample of the dataset (in Arabic Language), the first column is the comment ID and the second column is the comment text. Comment with id ends with 777bf is clear text and the comment id ends with c4d57 is identity hate comment (toxic comment)

Figure 4 shows the collected dataset. All the comments are classified into six different varieties. The comments are toxic, severe toxic, obscene, threat, insult, and identity hate. From that, we can find out the toxic comments and negative comments. By using these we can find the negative comments created by the bot. [27] The above plot shows the classification of negative comments present in the sample dataset. The toxic comments 15294 times occur in social media. The obscene comments occur in 8449 times. The insult comments occur in 7877 times. The 1595 times severe toxic

comment occurs. The 478 times threat comment occurs. The identity hate comment number of occurrence value is 1405. The toxic has the highest occurrence value. The comment classes are denoted by various colors.

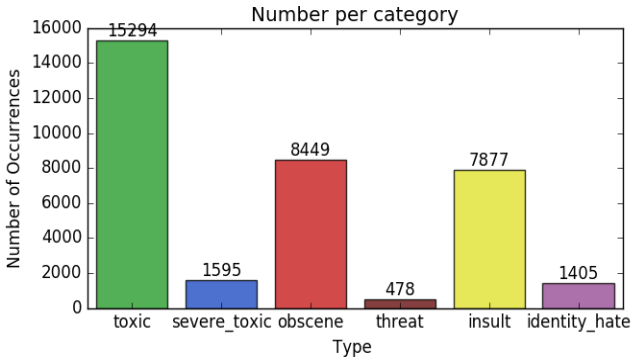

Figure 4 Different kinds of comments are classified into six different classifications based on the number of occurrences.

The correlation matrix, Figure 5 shows the relation between the different classification features. The black color stands for minimum correlation and the white color represents higher correlation. In the given figure, calculating the correlation matrix between the train dataset variables. The high correlation is obscene and insulting. There is two medium correlation shown in the figure such as toxic and insult, and toxic and obscene. The low correlation is severe toxic and insult, severe toxic and obscene. The correlation value starts at 0.15 and end at 0.90. The different colors are based on the correlation values. The lowest value color is black. The highest value color is white

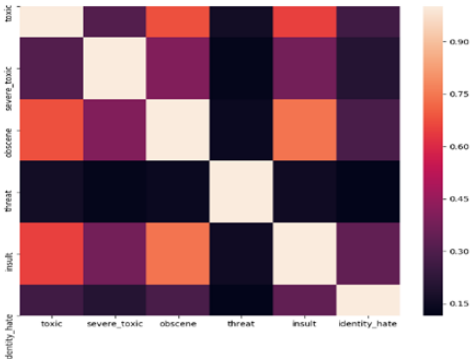

Figure 5. Correlation matrix shows the correlation between class features. Black squares show low correlation and white squares show high correlation.

B. Learning Phase

In Figure 6, classification accuracy score is shown to be 0.99. It displays the learning rate, accuracy score for training and validation. There are six learning rates as shown in the figure, they are 0.05, 0.1, 0.25, 0.5, 0.75 and 1. The x-axis denotes the learning rate and y-axis denotes the accuracy score for the training and validation. For learning rate, 1 training accuracy score value is 0.995 and the validation value is 0.996. The 0.75 learning rate the accuracy score validation and training are 0.995. The 0.5 and 0.25 learning rate has the accuracy score training is 0.995 and validation is 0.996. The 0.1 accuracy score validation is 0.582 and training is 0.589. The 0.05 learning rate accuracy score training value is 0.411 and validation value is 0.418.

C. The Classification Phase

The receiver operating characteristic (ROC) curve shows the true positive versus false positive rates. The x-axis denotes the false positive rate. The y-axis denotes the true positive rate. ROC curves are shown in Figure 7. The area under the ROC curve measures the accuracy of the classification model. It orders the test comments according to their class in decreasing order. The one at the top is the one that most likely belongs to positive class. Here, server toxic and obscene curves are most likely belonging to the positive classes. ROC curve accuracy is reduced if it is located near the diagonal. i.e. it is near 0.5.

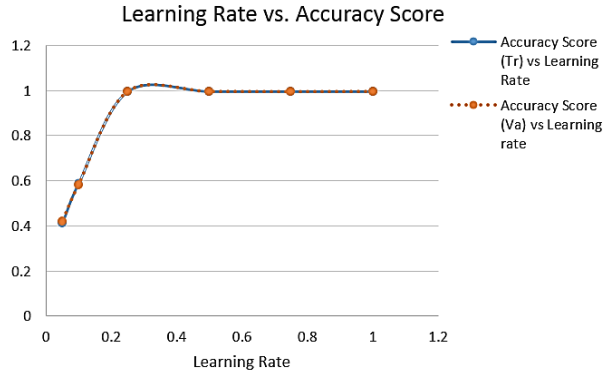

Figure 6. Learning rate and accuracy score for training and validation

The figure shows various negative comments for detecting the toxic comments. There are six different negative comments ROC area shown. Serve Toxic and obscene has the highest ROC area. That area value is 0.97. The threat, insult, and identity hate have the same ROC area value 0.96. The toxic ROC area value is 0.95. Those areas are denoted in various colors. The true negative value is 17689. The false positive value is 0. The false-negative value is 0. The true positive value is 12817.

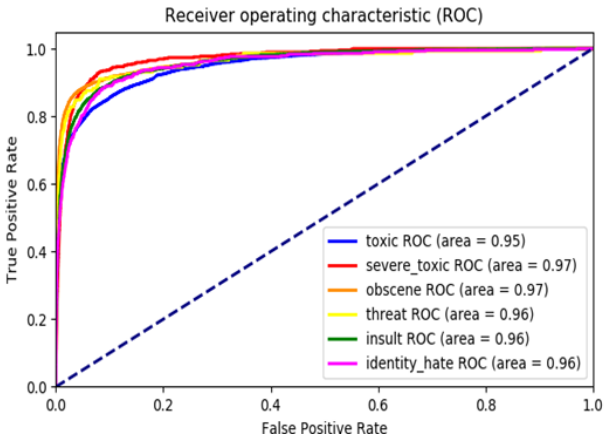

Figure 7. ROC curve, the figure presents ROC curve for toxic, sever toxic, obscene, threat, insult and identity hate features.

In our experiment, we defined three classes -1, 0 and 1. -1 defines the negative (bad) comments and 0 defines clean comments (Clean=Negative) and 1 refers to the undefined or unrecognized entries. For -1 class the precision, recall, and F_score are all 0.98. The support value is 17524. Class 0 has 99% accuracy and its corresponding recall and F_score value are 0.98. The support is 12812. Class 1 has the 0.00 precision, recall and F_score. The support is 123. There are two types of

averages denoted in the figure such as macro average and weighted average. The precision macro avg is 0.66. The recall and F_score have the 0.67 macro avg. The precision and F_score have the 0.98 weighted avg. the recall has the 0.99 weighted avg.

From the obtained results it is clear that the proposed model is comparatively less complex as well as more accurate than the other methods. In the ML-based toxic comments detection model, classification algorithm plays a significant role. It influences the performance of the algorithm. Most generally ID3, C4.5, KNN and Naive Bayes algorithms are used for the classification process. Also, SVM and ANN algorithms are used in some models. Each algorithm has its advantages and disadvantages. [2] Selection of the algorithm impacts on the final accuracy and time required for training and searching etc. Most of the ML-based machine learning algorithms use Naive Bayes and SVM algorithms. Among them, the Naive Bayes algorithm is very simple. But it doesn't provide higher accuracy. SVM algorithm provides higher accuracy but it requires more time for training. Especially in the case of large data size, it consumes more time. Similar to the SVM algorithm also requires more time. C4.5 and ID3 Algorithms gives noisy outputs. These are the major limitations only. The proposed model uses XGBoosting algorithm for the classification process [28].

In Figure 8, the classification accuracy of the existing methods as well as the proposed method is compared. The accuracy of the toxic comments identification model mainly depends on the classification algorithm used. the most accurate method is the proposed XGBoosting algorithm. It gives the accuracy value of 99.54. It is far more accurate than all the existing techniques.

It is noticed too that the nearest algorithm that performed comparative performance is the Random Forest. This is because XGBoosting is built by using decision trees which is the similar technique used in random forest.

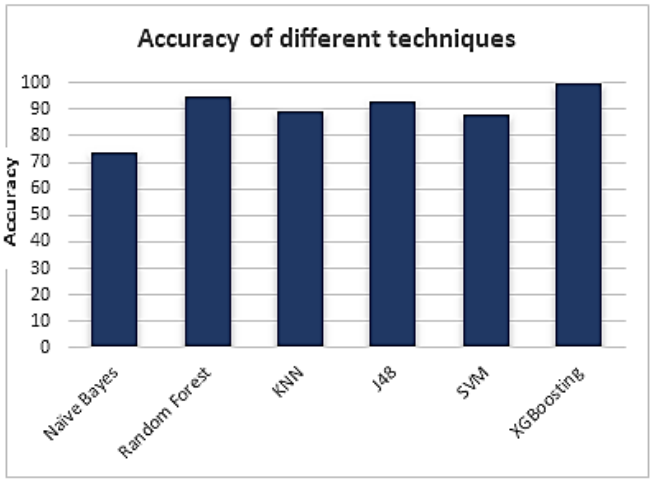

Figure 8. Performance comparisons for existing system and proposed system.

The feature importance is shown in figure 9, x-label denotes the f score and y-label denotes the features. There are six features are used for detecting the toxic comments. The features are threat, obscene, toxic, servere_toxic, insult and identity_hate. All features are differentiated by the various colors. The f score value starts from 0.00 and end at 0.30. The identity_hate has the highest F-score. Its value is 0.29781611. The servere_toxic has the lowest F-score value is 0.00862287. The insult f score value is 0.29191462. The toxic has the 0.20330871 f score. The obscene has the 0.09875891 f score. The threat has the 0.09957878 f score.

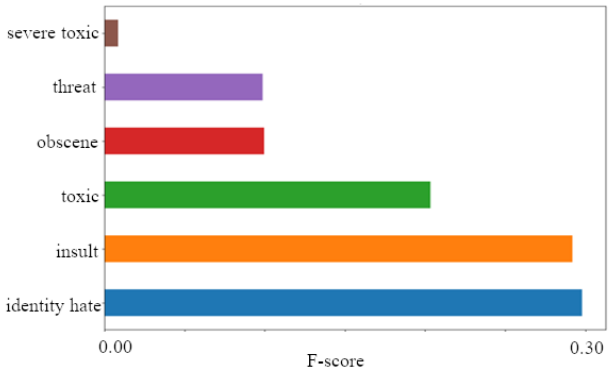

Figure 9. Feature importance, showing the interesting features ordered in ascending order. The most interesting feature is identity hate, the lowest interesting feature is the severe toxic feature.

V. Conclusion

Social media is adopted by companies and individuals for business and social activities. Attackers may attack companies or individual by publishing toxic comments about the established organization or famous person. Bots are created for such purpose. To identify these toxic comment bots, we have used XGBoost algorithm. XGBoost algorithm can successfully classify to toxic comments into the following categories, toxic, severe toxic, obscene, threat, insult, and identity hate. XGBoost was found to be more efficient in classification when compared to other algorithms such as SVM, KNN and Decision Trees. The XGBoost algorithm accuracy reaches more than 98%. Our future research may include the boosting algorithms into Intrusion Detection Systems (IDS) and Intrusion prevention Systems (IPS). IDS and IPS will help reduce such attacks and keep organizations and individuals on social media from being extruded and insulted from unknown attackers.

References

[1] Nonita Sharma, “XGBoost. The Extreme Gradient Boosting for Mining Applications”, *Munich, GRIN Verlag*, https://www.grin.com/document/415839 (2017)

[2] M. TOKMAK and E. KÜÇÜKSİLLE, "Detection of Windows Executable Malware Files with Deep Learning", *Bilge International Journal of Science and Technology Research*, 2019. Available: 10.30516/bilgesci.531801.

[3] T. Kumar, S. Sharma, H. Goel, S. Chaudhary and P. Jain, "A Novel Machine Learning Approach for Malware Detection", *SSRN Electronic Journal*, 2019. Available: 10.2139/ssrn.3383953.

[4] J. Bai and J. Wang, "Improving malware detection using multi-view ensemble learning", *Security and Communication Networks*, vol. 9, no. 17, pp. 4227-4241, 2016. Available: 10.1002/sec.1600.

[5] M. Eskandari and H. Raesi, "Frequent sub-graph mining for intelligent malware detection", *Security and*

Communication Networks, vol. 7, no. 11, pp. 1872-1886, 2014. Available: 10.1002/sec.902.

[6] Chen, Tianqi, and Carlos Guestrin. "XGboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016

[7] Y. Cho, "The Malware Detection Using Deep Learning based R-CNN", *Journal of Digital Contents Society*, vol. 19, no. 6, pp. 1177-1183, 2018. Available: 10.9728/dcs.2018.19.6.1177.

[8] M. Esmaeilpour and S. Mohammadkhani, "A new method for behavioural-based malware detection using reinforcement learning", *International Journal of Data Mining, Modelling and Management*, vol. 10, no. 4, p. 314, 2018.

[9] E. Karbab, M. Debbabi, A. Derhab and D. Mouheb, "MalDozer: Automatic framework for android malware detection using deep learning", *Digital Investigation*, vol. 24, pp. S48-S59, 2018. Available: 10.1016/j.diin.2018.01.007.

[10] S. Kaur and A. Kaur, "Detection of Malware of Code Clone using String Pattern Back Propagation Neural Network Algorithm", *Indian Journal of Science and Technology*, vol. 9, no. 33, 2016. Available: 10.17485/ijst/2016/v9i33/95880.

[11] A. Pektaş and T. Acarman, "Deep learning for effective Android malware detection using API call graph embeddings", *Soft Computing*, 2019. Available: 10.1007/s00500-019-03940-5.

[12] B. Khammas, "Malware Detection using Sub-Signatures and Machine Learning Technique", *Journal of Information Security Research*, vol. 9, no. 3, p. 96, 2018. Available: 10.6025/jisr/2018/9/3/96-106.

[13] Yang, Zhaocheng, and Rodrigo C. de Lamare. "Study of Sparsity-Aware Reduced-Dimension Beam-Doppler Space-Time Adaptive Processing." *arXiv preprint* arXiv:1903.01625 (2019).

[14] Mohammad, Fahim. "Is preprocessing of text really worth your time for online comment classification?." *arXiv preprint* arXiv:1806.02908 (2018).

[15] H. Hamdan, P. Bellot, and F. Bechet, "Lsislif : CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis," *no. SemEval*, pp. 753–758, 2015.

[16] S. Wang and C. Manning, "Baselines and Bigrams: Simple, Good Sentiment and Topic Classification," *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, no. July, pp. 90–94, 2012.

[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics, 5, 135-146* 2006.

[18] A. M.A. and J. C.D., "Automated multi-level malware detection system based on reconstructed semantic view of executables using machine learning techniques at VMM*", Future Generation Computer Systems*, vol. 79, pp. 431-446, 2018. Available: 10.1016/j.future.2017.06.002.

[19] MOUNIK, A., & KUMAR, D. P. "Malware Detection in Web Application using Content Integrity Verification", *International Journal of Recent Trends in Engineering and Research*, vol. 4, no. 3, pp. 460-464, 2018. Available: 10.23883/ijrter.2018.4151.ylppx.

[20] M. Narouei, M. Ahmadi, G. Giacinto, H. Takabi and A. Sami, "DLLMiner: structural mining for malware detection", *Security and Communication Networks*, vol. 8, no. 18, pp. 3311-3322, 2015. Available: 10.1002/sec.1255.

[21] C. Ravi and R. Manoharan, "Malware Detection using Windows API Sequence and Machine Learning", *International Journal of Computer Applications*, vol. 43, no. 17, pp. 12-16, 2012. Available: 10.5120/6194-8715.

[22] Natekin A and Knoll A "Gradient boosting machines", a tutorial. *Front. Neurorobot*. doi: 10.3389/fnbot.2013.00021, 2013

[23] P. Singhal, "Malware Detection Module using Machine Learning Algorithms to Assist in Centralized Security in Enterprise Networks", *International Journal of Network Security & Its Applications*, vol. 4, no. 1, pp. 61-67, 2012. Available: 10.5121/ijnsa.2012.4106.

[24] W. Zhong and F. Gu, "A multi-level deep learning system for malware detection", *Expert Systems with Applications*, vol. 133, pp. 151-162, 2019. Available: 10.1016/j.eswa.2019.04.064.

[25] R. RIASAT, M. SAKEENA, A. SADIQ, C. WANG, C. ZHANG and Y. WANG, "Machine Learning Approach for Malware Detection by Using APKs", *DEStech Transactions on Computer Science and Engineering*, no., 2017. Available: 10.12783/dtcse/cnsce2017/8883.Available: 10.1504/ijdmmm.2018.10015880.

[26] P. Sodhi, N. Awasthi and V. Sharma, "Introduction to Machine Learning and Its Basic Application in Python", *SSRN Electronic Journal*, 2019. Available: 10.2139/ssrn.3323796.

[27] P. Srivastava and M. Raj, "Feature extraction for enhanced malware detection using genetic algorithm", *International Journal of Engineering & Technology*, vol. 7, no. 28, p. 444, 2018. Available: 10.14419/ijet.v7i2.8.10479.

[28] Do Quan. "*Jigsaw Unintended Bias in Toxicity Classification*." Bachelor's thesis Electrical and Automation Engineering Spring 2019.

Author Biographies

Osama Hosameldeen (Osama Hosam) Is a research associate in SRTA-City, Alexandria, Egypt. In 2007 he received his MSc. In computer systems and engineering from Azhar University, He pursued his PhD study in Hunan University, China and worked in parallel in Nanjing University of Technology; in 2011 he received his PhD in Computer Science and Engineering. In 2013 he worked as an Assistant Professor in at the Collage of Computer Science and Engineering in Yanbu. In 2017 he is promoted to be an Associate Professor in the field of Computer and Information Security.