Developing a Protective – Preventive and Machine Learning Based Model on Child Abuse

Fatih Mert, Muhammed Ali Aydin and Abdül Halim Zaim

February 3, 2021

Developing a Protective – Preventive and Machine Learning Based Model on Child Abuse

Fatih MERT
Department of Computer Engineering
Istanbul Commerce University
Istanbul, Turkey
fatih.mert@istanbulticaret.edu.tr

Muhammed Ali AYDIN
Department of Computer Engineering
Istanbul University - Cerrahpasa
Istanbul, Turkey
aydinali@istanbul.edu.tr

Abdül Halim ZAIM
Department of Computer Engineering
Istanbul Commerce University
Istanbul, Turkey
azaim@ticaret.edu.tr

***Abstract* — Online grooming is an ever-increasing problem in societies and the time spent online is recently started to rise drastically. People can become anonymous whilst posting, sharing his/her own opinion, and being a part of online chatting. Option to be anonymous also brings together the chance for hiding personal identity when making an attempt on illegal activities. Online grooming is one of the significant areas of aforementioned actions and sexual predators can easily use online chatting platforms to quickly build a friendly relationship with children or teenagers to gain their trust and make them share their obscene media files. These sexual predators mostly try to convince their victims to meet and it may lead to having sexual intercourse with a minor. In order to draw attention to the huge challenge that most societies face, this study mainly aims to identify predators in the early stage of online communication. The objective is to do an investigation to detect child grooming through online chat records by using Machine Learning techniques. In the first part of the study, it has been achieved to make a multi-label classification on a Wikipedia dataset with more than 97 percent accuracy, where a given text gets classified based on the toxicity types. The outcome of this work is also used in the second stage and herein PAN12 dataset has been used to train and test our model. We have ended up with more than 92 percent accuracy, where suspicious conversation messages from the chat records get identified and sexual predators can be recognized.**

Keywords— Child Abuse Detection, Online Sexual Predator Identification, Multi-Label Text Classification, Machine Learning

I. INTRODUCTION

As of October 2020, the number of active internet users has been reached around 4.66 billion people all over the world, amounting to nearly 50 percent of the whole population globally [1]. The number of messaging application users also exceeds billions world-wide. As people tend to use social applications with the spread of internet usage, it comes with unwanted troubles as well. More than 80 percent of the youth who resided in the USA could reach the internet and children whose ages were between 5 – 16 were spending nearly 7 hours a day on the devices having internet access. Even though the internet is a fabulous source of information, it may also become an environment full of danger, especially for children. One of the major reasons behind this issue is that there is no recognized way to regulate the usage of the internet. People can become anonymous whilst posting, sharing his/her own opinion, and being a part of online communication. Option to be anonymous also brings together the chance for hiding personal identity when making an attempt on illegal activities. Hence, any malevolent person can easily attempt to the solicitation, both in virtual and real life. Online solicitation addresses the moment when an adult asks for having sexual intercourse, being a part of undesired sexual actions, or having sexual talks in the online areas. When youth are compared to adults, their level of sense for making an inference regarding having an inkling of potential threats waiting after interaction with people who have ill-will against themselves. Based on the outcomes of a study, nearly 20 percent of the youths have been subjugated to sexual content without their acquiescence and nearly 10 percent of them have experienced unwelcome online sexual abuse [2]. Herein, we should not overlook the unreported solicitations, since most of the children feel quite ashamed and guilty preventing them to explicitly declare the situation they have been going through. Moreover, they may even not be aware of that the fact that they were abused. Online-facilitated child abuse could be done through many ways: The production, dissemination or possession of CSAM (Child Sexual Abuse Materials), also knowns as ‘’child pornography’’ in the general acceptance, sexting (sending or receiving of sexual texts or media files such as pictures or videos through technology usage), revenge pornography, online child grooming (befriending and building an emotional bridge with children to heighten their exiting curiosity about sex, with the ultimate aim of meeting them in real, by considering sexual benefits), active sexual harassment, sexual extortion (also known as sextortion), abuse of children over online prostitution, live streaming of sexual incident, and etc. [3] Online grooming is one of the significant ways of aforementioned sexual abuse actions and sexual predators can easily use online chatting platforms to quickly build a friendly relationship with children or teenagers to gain their trust and make them share their obscene media files. These sexual predators mostly try to convince their victims to meet and it may lead to having sexual intercourse with a minor. In order to draw attention to the huge challenge that most societies face, this study mainly aims to identify predators in the early stage of online communication. The objective is to do an investigation to detect child grooming through online chat records by using Machine Learning techniques.

Structure of the rest of this paper is given below:

Section 2, introduces our project and mainly gives the background with the basic understanding and explanations of online child abuse. A detailed summary of the related work conducted in the literature is explained throughout this chapter. Section 3, describes the methodology that has been used throughout this study. Section 4, gives a presentation for the results of the conducted research and whole study. Section 5, concludes with the suggestions and describes possible future works.

II. LITERATURE REVIEW

A. Related Work

A study focusing on child abuse identification in the public health sector with the examination of medical records, used feature extraction from the word clouds, with the help of classifiers such as SVM. The overall performance of this study has been stated as good-enough for daily usage. [4]

Another study conducted in the collaboration with the Swedish Financial Coalition targeted to make a classification for illegal advertisement on Dark Net. In order to perform the algorithm by evaluating several classification models and feature extraction techniques, deep learning was used and it was seen that these deep learning models outperformed the standard methods. [5]

State-of-the-art technologies were presented for analyzing internet crimes against children in a study, where the main purpose was to protect children from being abused by online predators, by developing automated tools. As a result, a program was developed, helping to correctly identify the online sexual predators 60 percent of the time. Following several updates in the second experiment, the identification ratio has been reached 93 percent. [6]

In another study, each line in a conversation has been labeled and communication theories with computer algorithms were used for the identification of predatory messages. After using different machine learning algorithms that classified the lines based on a rule-based approach and phrase matching, the approach labeled the lines with 83.11 percent accuracy where the experiment included 33 unique conversations. [7]

The last sample work in the literature provided an overview of the PAN12 competition focused on sexual predator identification task internationally. This contest was considered a combination of two sub-challenges, the first, being the challenge to identify possible whole predators from the given chat logs, which consisted of both predatory and non-predatory data within. The latter challenge was the task to make an identification for which of the predators' lines in the conversations can be marked as a moment for abusive behaviors to take place. [8]

B. Types of Machine Learning Models

Below figure gives an overview for the categorization of machine learning algorithms and both of our tasks fall into the category of supervised learning.

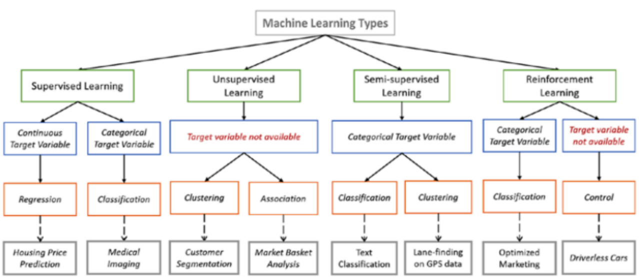


Fig. 1. Types of machine learning algorithms [9]

1. Supervised Learning

Given a set of data points as $\{x^{(1)}, \ldots, x^{(m)}\}$ and this dataset is associated with a set of outputs as $\{y^{(1)}, \ldots, y^{(m)}\}$. Then we would like to come up with a classifier which will learn how to make prediction (predict y from x). In our work, we used 5 different supervised learning models in order to implement binary and multi-label classification tasks.

a) Naïve Bayes

Naive Bayes is a probabilistic machine learning model that is used for classification tasks. It is supposed that the features of each data point will be independent of each other. It is mostly used in sentiment analysis, recommendation systems, and spam filtering tasks. It is also very widely used for text classification tasks. The advantage is speed and being easy-to-use. However, there is also a disadvantage related to the requirement of predictors to be independent. Typical types of the Naive Bayes classifiers are as follows: Multinomial Naive Bayes, Bernoulli Naive Bayes, Gaussian Naive Bayes, and so on.

b) Support Vector Machine (SVM)

Support Vector Machine (SVM) is a type of classifier which is described by a hyperplane which is a V-1 dimensional subspace of a V-dimensional vector space (Christopher M.). The main goal is to find the line which will maximize the minimum distance to the line.

With the use of kernels, SVM is more powerful and helpful for solving classification problems. Some of the kernel options are Linear, Sigmoid, Gaussian, Polynomial... If the Linear kernel is used, learning of the hyperplane gets performed by making the problem transform into a linear algebra problem. SVM Classifier has a regularization parameter called C, in order to detect how much of the misclassification is tolerated for each and every data given as an input. Another parameter of SVM is gamma and it describes how far the influence of a particular input from training reaches. Majority of the time SVMs are chosen for classification tasks, especially binary classification.

c) Logistic Regression

Logistic Regression is one of the most commonly used methods for solving classification problems. This model works for computing the logarithm of the odds as a linear combination of predictors (independent variables). Logistic Regression is mainly a combination of the Sigmoid function and linear regression equation. The advantageous aspect of logistic regression is that high computational power is not needed. It is very easy to use and mostly used by data scientists. In contrast, there is a disadvantage of not being able to deal with a big number of features, and this classifier is not powerful when it comes to overfitting.

d) K-Nearest Neighbors (K-NN)

The K-Nearest Neighborhood (KNN) algorithm is one of the easy-to-implement supervised learning algorithms. It is used in the solution of both classification and regression problems but mostly used in the solution of classification problems in the industry. First, the k parameter is determined. This parameter is the number of neighbors closest to a given point. For example: Let k = 2. In this case, the classification will be made according to the closest 2 neighbors. With the help of the relevant distance functions, the distance of the new data to be included in the sample data set is calculated one by one according to the existing data. The nearest neighbors from the relevant distances are considered. It is assigned to the class of k neighbors or neighbors according to the attribute values.

The selected class is considered to be the class of the observation value expected to be estimated. In other words, the new data is labeled.

e) AdaBoost

AdaBoost, in other words, Adaptive Boosting, is a commonly used machine learning method and it is known as one of the boosting algorithms. Boosting algorithms are used as a collection of classifiers with low accuracy, in order to build a highly accurate classifier. Boosting algorithms are not that much affected by the problem of overfitting. AdaBoost, Gradient Tree Boosting, and XGBoost are the most commonly used boosting algorithms and in this study we used AdaBoost. The main logic behind AdaBoost is about setting the classifier weights and sampled training data in each and every iteration. That way we can make sure of the accuracy of unusual records.

2. Unsupervised Learning

The main goal of unsupervised learning is to find the hidden layers in unlabeled data, $\{x^{(1)}, \ldots, x^{(m)}\}$. Here, the algorithm tries to identify patterns by studying the data. Unlike supervised learning, the machine makes a determination regarding the correlation and relationships checking the available data. The task for making the dataset convert into an organized version, the machine groups the data into clusters so that it will look more organized.

3. Semi-supervised Learning

It is quite similar to supervised learning. However, it combines the work on both unlabeled and labeled datasets. That way, the machine learns how to label the unlabeled data.

4. Reinforcement Learning

The main focus is to provide a set of actions and processes that can be considered as regimented learning. After monitoring and evaluating each and every result for the aim of determining the optimal one, this learning type defines a set of rules in the beginning. In this approach, the machine is taught by trial and error. By learning from the previous experiences, the algorithm adopts as a response to the situation and tries to get the possibly best result.

C. Evaluation Metrics

Loss Function: It is defined as a function that takes the predicted values of z to correspond to the real value of y as input and shows how different they are. Some of the most commonly used loss functions are least-square error, logistic loss, hinge loss, cross-entropy, hamming loss and etc.

$$L : (z, y) \in R \times Y \longmapsto L(z, y) \in R \qquad (1)$$

Confusion Matrix: It is used in order to have a complete representation for the model performance assessment. The figure showing a simple confusion matrix is given as below:

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positive	False Negative
	Negative	False Positive	True Negative

Fig. 2. Binary confusion matrix

Accuracy Score: It is the proportion of correctly classified predictions over the total number of predictions.

Precision Score: It is the proportion of correctly predicted inputs over the total number of samples which belongs to that particular class.

Recall Score: It is the proportion of correctly predicted inputs given all existing samples of that class.

F- Score: It refers to the harmonic mean of Precision and Recall scores.

TABLE I. MOST COMMONLY USED EVALUATION METRICS

Evaluation Metric	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FN}$
F1-Score	$\frac{2TP}{2TP + FP + FN}$
True Positive Rate	$\frac{TP}{TP + FN}$

Receiver Operating Curve (ROC: It refers to the plot representation of True Positive Rate (TPR) with respect to False Positive Rate (FPR).

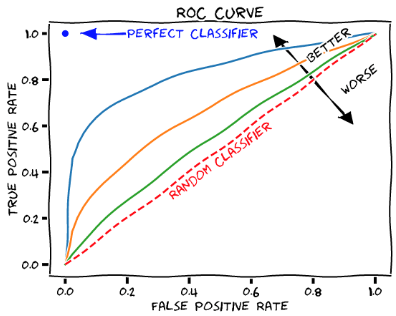


Fig. 3. Evaluation of ROC curves [10]

Precision-Recall Curve: It is the summarization of the trade-off between the True Positive Rate (TPR) and positive predicted value. It is more useful for imbalanced datasets.

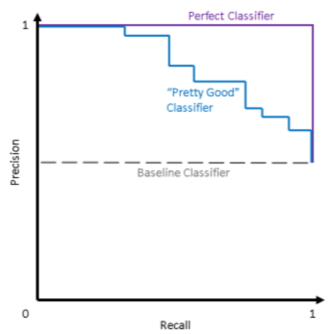


Fig. 4. Evaluation of Precision-Recall curves [11]

K-Fold Cross Validation: It is a widely used method for performance assessment. When the data is scarce, it is most of the time helpful to split the dataset several times creating multiple validations, as well as multiple training and test sets for making the assessment. A sample representation of K-Fold cross-validation is shown in Fig. 5.

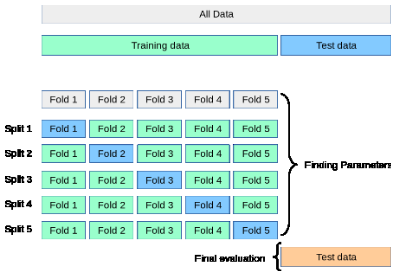


Fig. 5. K-Fold cross validation [12]

III. METHODOLOGY

A. Data Gathering

Wikipedia Comment Dataset: This dataset is provided with a large number of Wikipedia comments that have been labeled by human raters for examining toxic behaviors. Dataset has been obtained via kaggle.com.

PAN12 Dataset: This dataset contains the training and test corpus for the ‘’Sexual Predator Identification’’ task of the Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN) Lab. Dataset has been obtained via zenodo.org.

B. Pseudocode for Toxic_Comment_Classifier

```
1.df_train, df_test: Dataframes for train and test data
2.label_list: [toxic, severe_toxic, obscene, threat, insult, identity_hate]
3.contractions: List of abbreviations for normal and abusive languages, as well as encrypted sexting speech
4.test_size: 0.33
5.classification_models:[MultinomialNB, LinearSVC, LogisticRegression, K-NN, AdaBoost]
6.for each comment_text in df_train and df_test:
        #Apply cleaning
        6.1.Remove HTML tags
        6.2.Remove punctuations
        6.3.Remove non-alphanumeric characters
        6.4.Expand contractions
        6.5.Apply stop-words removal
        6.6.Apply stemming
        6.7.Remove most commonly used words
        6.8.Remove most rarely used words
7.Apply oversampling to deal with data imbalance
8.train,test: Split df_train into 2 dataframes with proportional to test_size
9.for each model in classification_models:
        9.1.X_train: vectorized train
        9.2.X_test: vectorized test
        9.3.ngram_range: (1,2)
        9.4.for each label in label_list:
                9.4.1.model.fit(X_train, train[label])
                9.4.2.prediction: model.predict(X_test)
        9.5.Display performance evaluation results comparing
        prediction and test[label]
```

C. Pseudocode for Sexual_Predator_Identifier

```
1.train_data, test_data: get raw data in XML format
2.for each conversation in train_data, test_data:
        2.1.if number_of_authors == 1:
                2.1.1. remove conversation
        2.2.if number_of_messages < 5:
                2.2.1 remove conversation
        2.3.if ratio_of_unrecognized_chars > 0.65
                2.3.1 remove conversation
3.for each XML tag in train_data:
        3.1. convert tag into list
4.df_train, df_test: Dataframes converted from tags for train and test data
5.label_list: [sexual_predator]
6.contractions: List of abbreviations for normal and abusive languages, as well as encrypted sexting speech
7.test_size: 0.20
8.classification_models:[MultinomialNB, LinearSVC, LogisticRegression, K-NN, AdaBoost]
9.for each chat_message in df_train and df_test:
        #Apply cleaning
        9.1.Remove chat_message if it has 1 word
        9.2.Remove HTML tags
        9.3.Remove punctuations
        9.4.Remove non-alphanumeric characters
        9.5.Expand contractions
        9.6.Apply stop-words removal
        9.7.Apply stemming
        9.8.Remove most commonly used words
        9.9.Remove most rarely used words
        9.10. Apply spell-check
10.Apply oversampling to deal with data imbalance
11.Identify abusive chat messages using previously built Toxic_Comment_Classifier and mark them in the newly added column called abusive_message
12.X_train,X_test, y_train, y_test: Split df_train using test_size
13.for each model in classification_models:
        13.1.X_train: TF-IDF vectorized train
        13.2.X_test: TF-IDF vectorized test
        13.3.ngram_range: (1,2)
        13.4.model.fit(X_train, y_train)
        13.5.prediction: model.predict(X_test)
        13.6.Apply hyper-parameter tuning to find best
        parameters
14.Display performance evaluation results comparing prediction and y_test
```

IV. RESEARCH FINDINGS & DISCUSSION

In the first stage of our work, we have concentrated on classification of toxic comments, while in the second stage we focused on the task of sexual predator identification. For both of the specified sub-tasks, we basically used ROC curves and Precision-Recall curves, in order to measure and compare the performance results. Since our datasets are highly imbalanced, meaning that the distribution of the labels are not homogeneous, we could not rely on the Accuracy score. Most of the time, it needs to be clearly defined which classification metrics should be chosen, in the light of the problem domain and priorities. For example; we can better choose depending on what we would like to predict (class labels or probabilities). Assume that we want to predict the probabilities and we need the class labels; Precision-Recall curves would be more useful if the positive class label is more important for us, whereas ROC curves would be more useful if both of the labels are equally important. In our case checked both of them. Depending on further scenarios, both metrics would be giving an idea about which classification model should be used. For the purpose of a clearer interpretation, we also compared the basic metrics (Accuracy, F-Score, Precision, Recall) and displayed the graph considering their average values.

A. *Results of Toxic Comment Classification*

ROC Curves:

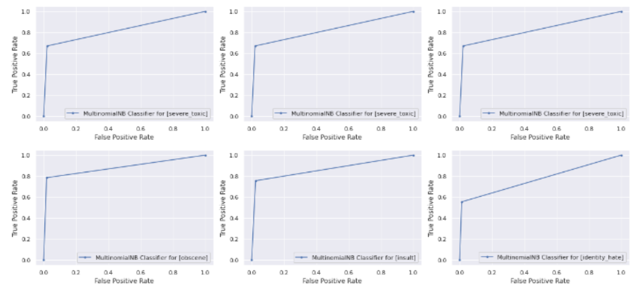


Fig. 6. ROC curves of Multinomial Naïve Bayes

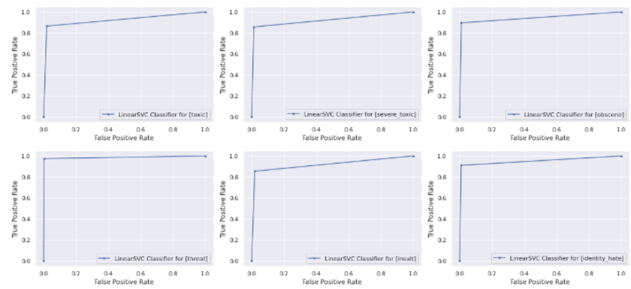


Fig. 7. ROC curves of Linear SVC

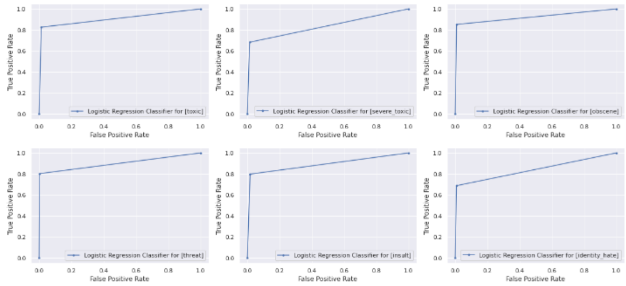


Fig. 8. ROC curves of Logistic Regression

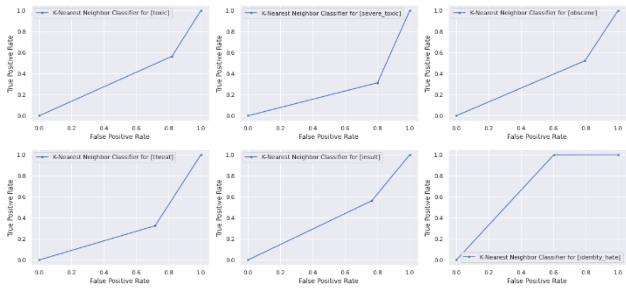


Fig. 9. ROC curves of K-NN

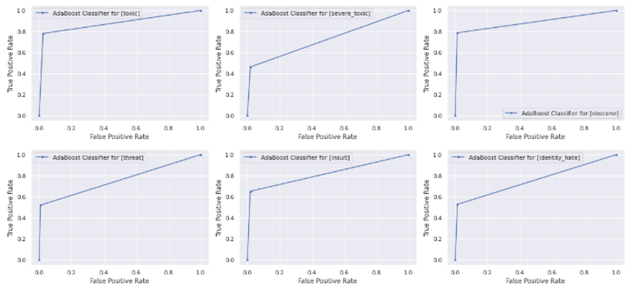


Fig. 10. ROC curves of AdaBoost

ROC curves are basically drawn using False Positive Rates and True Positive Rates. As we checked the ROC curves of our classifier models, we have seen that LinearSVC gives the best results, since the curves for all of the 6 labels are closest to the perfect curve given in Figure 3. We can also see that all of our classifier models performed better than normal case (random classifier) but K-NN classifier. As we checked the literature, it was seen that K-NN was not highly preferred as the others, as well.

Precision - Recall Curves:

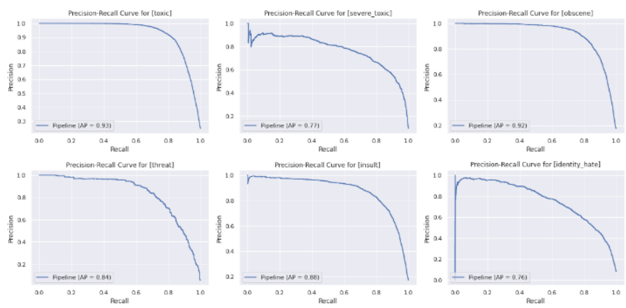


Fig. 11. Precision-Recall curves of Multinomial Naïve Bayes

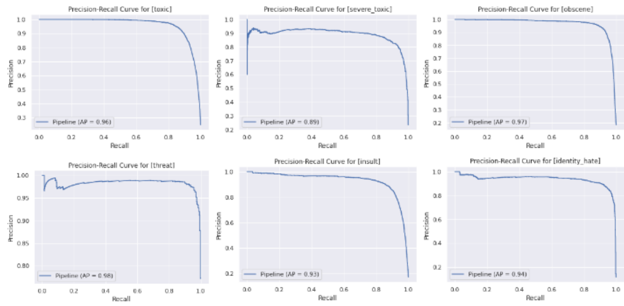


Fig. 12. Precision-Recall curves of Linear SVC

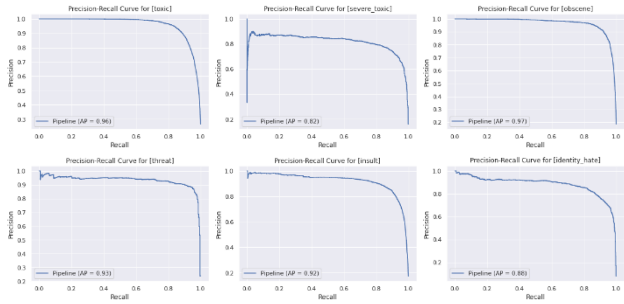


Fig. 13. Precision-Recall curves of Logistic Regression

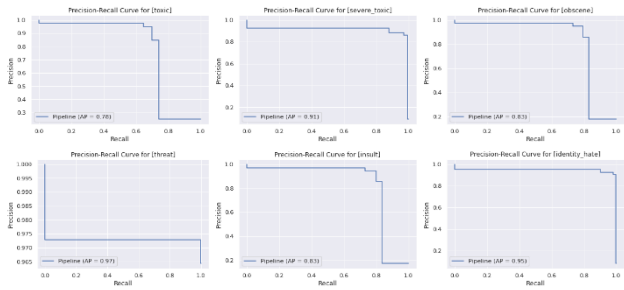


Fig. 14. Precision-Recall curves of K-Nearest Neighbors

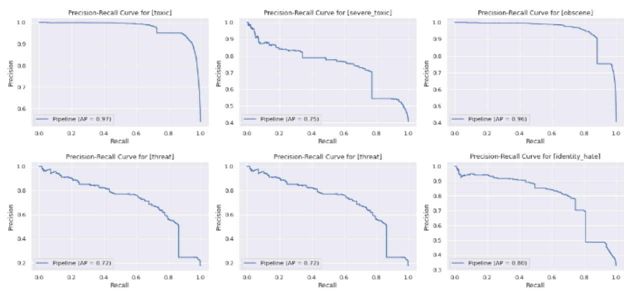
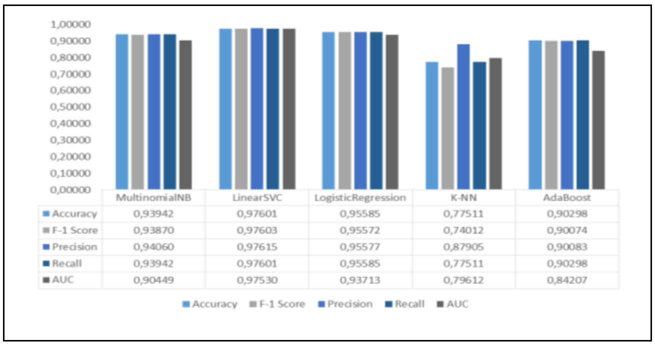


Fig. 15. Precision-Recall curves of AdaBoost

Precision-Recall curves are used to depict the relationship between precision (also known as the positive predicted value) and recall (also knowns as sensitivity). These curves often tend to have frequent zigzags with up and downs in the shape and for that reason, if we combine several classifiers in a single representation, it is more possible for them to intercept with each other when compared to ROC Curves. In our case, the Precision-Recall curves of the Linear SVC model are the best outputs, since their shape is closest to the perfect representation given in Fig 4. We can also see that Multinomial Naïve Bayes and Logistic Regression models have performed quite well.



	MultinomialNB	LinearSVC	LogisticRegression	K-NN	AdaBoost
Accuracy	0,93942	0,97601	0,95585	0,77511	0,90298
F-1 Score	0,93870	0,97603	0,95572	0,74012	0,90074
Precision	0,94060	0,97615	0,95577	0,87905	0,90083
Recall	0,93942	0,97601	0,95585	0,77511	0,90298
AUC	0,90449	0,97530	0,93713	0,79612	0,84207

Fig. 16. Performance metrics comparison of classifier models

We can also compare the overall performance results in the light of the table given in Fig 16. For each metric, it is clear that Linear SVC gives the best results with more than 97 percent achievement. According to the table, Logistic Regression would be our second preferrence and it could be followed by Multinomial Naïve Bayes and AdaBoost. However, K-NN is not a good option to use as our classifier model.

B. *Results of Sexual Predator Identification*

While performing sexual predator identification task, we again used 5 different classification models. These are Multinomial Naïve Bayes, Linear SVC, Logistic Regression, K-NN and AdaBoost. Unlike the previous task, this time our target label was only sexual_predator. Since it is not a multi-label classification problem as before, the overall result was not similar to the toxic comment classification and this time K-NN outperformed the other models.

ROC Curves:

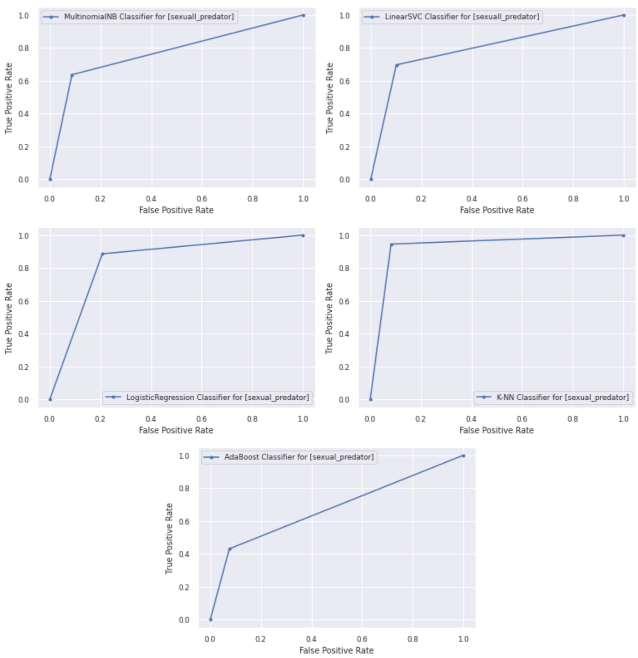


Fig. 17. ROC curves of classifier models

The ROC curves given above tells us that K-NN classifier outperforms the other classification model as its shape is much closer to the perfect case and the area under the curve is higher than the other models. We can also note that AdaBoost is not performing as expected. K-NN is followed by Logistic Regression, Linear SVC and Multinomial Naïve Bayes models as well.

Precision - Recall Curves:

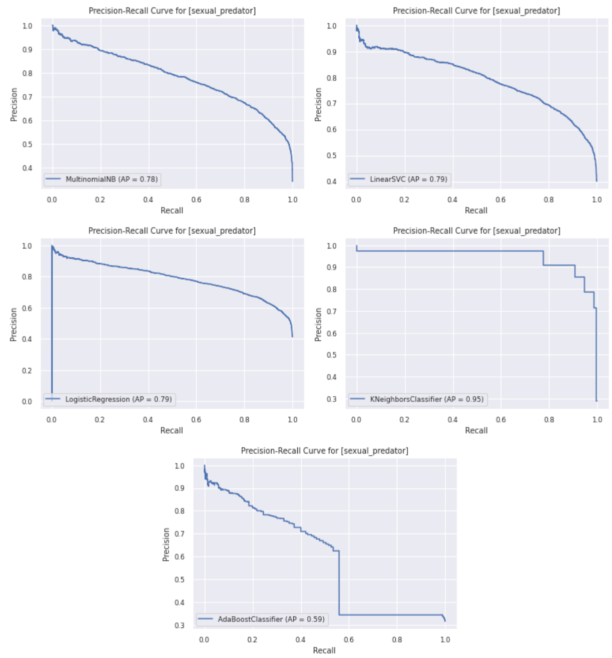
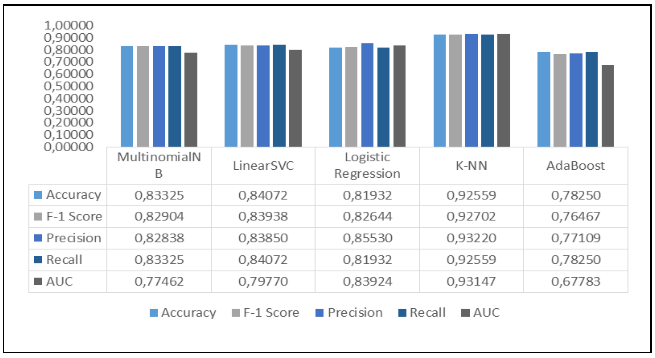


Fig. 18. Precision-Recall curves of classifier models

Based on the generated Precision-Recall curves, we can clearly see that K-NN model gives us the best results and it has the highest area under the curve. Multinomial Naïve Bayes, Linear SVC, Logistic Regression could be also considered as a good model, but AdaBoost did not give us a shape as expected, based on the perfecrt classification curve shown in Fig 4.



	MultinomialNB	LinearSVC	Logistic Regression	K-NN	AdaBoost
Accuracy	0,83325	0,84072	0,81932	0,92559	0,78250
F-1 Score	0,82904	0,83938	0,82644	0,92702	0,76467
Precision	0,82838	0,83850	0,85530	0,93220	0,77109
Recall	0,83325	0,84072	0,81932	0,92559	0,78250
AUC	0,77462	0,79770	0,83924	0,93147	0,67783

Fig. 19. Performance metrics comparison of classifier models

Based on our performance metric comparison table, we can see that our K-NN classification model outperformed the other models for all of the metrics given above (Accuracy, F-1 Score, Precision, Recall and AUC). It is also seen that Multinomial Naïve Bayes and Linear SVC models perfomed almost the same. However, AdaBoost classifier was a poor model when compared to others.

V. Conclusion

In light of the graphical representations of the most commonly used evaluation metrics, we can infer that Linear SVC is the most appropriate classifier model for our toxic comment classification task. When compared to the other ones, it is clear that Linear SVC has the best scores for Accuracy, Precision, Recall, F1, and AUC values. As for the sexual predator identification task, we can see that K-NN outperforms the other classifier models since it has the best results for the same metrics. Once we checked the training and prediction times, K-NN became the slowest one due to dealing with neighborhood selection, whereas Logistic Regression is the fastest model. We also calculated hamming loss scores for each and every classifier and the more accuracy we had for our model, the less loss score we observed. All of the drawn graphs are consistent with each other. For the ROC and Precision-Recall curves, we checked our results based on Fig 3 and Fig 4. We believe that we have successfully implemented and combined the concept of toxicity classification and sexual predator identification in our work, by generating our classifier models following a series of Natural Language Process tasks. As a result, we are able to identify sexual predators for a given set of conversations, as well as we can highlight the abusive messages with the help of our toxic comment classifier.

VI. FUTURE WORK

In the current work, we could not concentrate on image data, due to time limitations. Instead, we started with the text type of dataset. However, most of the Child Sexual Abuse Materials are images and media files. Hence, image processing algorithms could be inserted into our work for the betterment of the predator identification task. Then, a more in-depth version of pre-filtering and text processing activities could be done and new abbreviations that recently became popular but not unofficial yet could be discovered. As a more common trend in the literature, Deep Learning algorithms could be tried to come up with better results. Lastly, in the upcoming revisions, it would be helpful to focus on the Turkish language, since there is no sufficient number of academic research studies in the domain of preventing online child abuse through Machine Learning methodologies.

REFERENCES

[1] https://www.statista.com/statistics/265147/number-of-worldwide-internet-users-by-region/

[2] C. Azzopardi, R. Eirich, C. L. Rash, S. MacDonald, S. Medigan, A meta-analysis of the prevalence of child sexual abuse disclosure in forensic settings, 2018, Elsevier.

[3] https://www.ecpat.org/what-we-do/online-child-sexual-exploitation/

[4] C. Amrit, T. Paauw, R. Aly, M. Lavric, Identifying child abuse through text mining and machine learning, Expert Systems with Applications, 2017.

[5] H. Adamsson, Classification of illegal advertisement working with imbalanced class distributions using machine learning, DiVA, 2017.

[6] A. Kontosthathis, L. Edwards, A. Leatherman, Text mining and cybercrime, Wiley Online Library, 2010.

[7] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, E. Jakubowski, Perverted Justice: Learning to identify internet sexual predation, International Journal of Electronic Commerce 15, 3, 103-122, 2011.

[8] G. Inches, F. Crestani, Overview of the international sexual predator identification competition at PAN-2012, CLEF, 2012.

[9] https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861

[10] https://analyticsindiamag.com/beginners-guide-to-understanding-roc-curve-how-to-find-the-perfect-probability-threshold/

[11] https://medium.com/@douglaspsteen/precision-recall-curves-d32e5b290248

[12] https://scikit-learn.org/stable/modules/cross_validation.html

[13] C. Morris, G. Hirst, Identifying online sexual predators by SVM classification with lexical and behavioral features, CLEF, 2012.

[14] M. A. Saif, A. N. Medvedev, M. A. Medvedev, T. Atanasova, Classification of online toxic comments using the logistic regression and neural networks models, AIP Conference Proceedings, 2018.

[15] M. Ibrahim, M. Torki, N. El-Makky, Imbalanced toxic comments classification using data augmentation and deep learning, ICMLA, 2018.

[16] C. Cardei, T. Rebedea, Detecting sexual predators in chats using behavioural features and imbalanced learning, Cambridge University Press, 2015

[17] M. Ebrahimi, C. Y. Suen, O. Ormandijeva, A. Krzyzak, Recognizing predatory chat documents using semi-supervised anomaly detection, Society for Imaging Science and Technology, 2016.

[18] H. J. Escalante, E. Villatoro-Tello, A. Ju´arez, L. Villaseñor, M. Montes-y-G´omez, Sexual predator detection in chats with chained classifiers, Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013.

[19] J. I. Rodríguez, S. R. Durán, D. Díaz-López, J. Pastor-Galindo, F. G. Mármol, C3-Sex: A conversational agent to detect online sex offenders, Electronics, 2020.

[20] M. W. RahmanMiah, J. Yearwood, S. Kulkarni, Detection of child exploiting chats from a mixed chat dataset as a text classification task, Australasian Language Technology Association Workshop, 2011.