National College of Ireland

Cybercrime detection in communications: An Experimental case of cyber Sexual Harassment Accuracy detection on Twitter Using Supervised Learning Classifiers

MSc Internship
Cybersecurity

EZE KENNETH C.
Student ID: X19131178

School of Computing
National College of Ireland

Supervisor: Mr. Niall Heffernan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	EZE KENNETH C.
Student ID:	X19131178
Programme:	Cybersecurity **Year:** 2019/2020
Module:	Internship
Supervisor:	Mr. Niall Heffernan
Submission Due Date:	17th August 2020
Project Title:	Cybercrime detection in communications: An Experimental case of cyber Sexual harassment Accuracy detection on Twitter Using Supervised Learning Classifiers
Word Count:	5177 **Page Count** 17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.
I agree to an electronic copy of my thesis being made publicly available on NORMA the National College of Ireland's Institutional Repository for consultation.

Signature: Kenneth Eze

Date: 14th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Cybercrime detection in communications: An Experimental case of cyber Sexual harassment Accuracy detection on Twitter Using Supervised Learning Classifiers

EZE KENNETH C.
X19131178

Abstract

With the growth of social media platforms and how it fits easily in our daily routine, the issue of cyber sexual harassment which is a disturbing online misbehavior with troubling consequences has been an emerging problem in schools, home and work places.
Having to understand the implications of cyberbullying has become a major concern in our society. To assist in the investigation and preventive efforts, the introduction of systems automatically analysing acts of cyber sexual harassment with minimal false positives has become a common trend.
Automatic detection of possible harassment or threats can assist law enforcement agencies in making informed decisions, provide behavioral patterns of the aggressor or forensic evidence through these social platforms. And this in turn, can curb escalation of these dangers in real life scenarios and create a more secure online experience.
The research covers the use of supervised machine learning algorithms in the analysis of dataset obtained from digital media and form a comparison to determine the algorithm which gives the highest accuracy in detecting activity of sexual harassment online.
.

Keyword: *Machine learning, online communication, cybercrime, online social network, Twitter*

1 Introduction

As social network grows exponentially it has provided a lot of social interaction among various users in different locations and this causes a huge amount of user-generated communication data. In recent years, Cyber Sexual Harassment a form of victimization has grown to be a huge problem with the growth of online communication and social media.
Cyber victimization needs to be understood and addressed from various perspectives and automatic detection and prevention of these incidents can substantially help to tackle this problem.
A recent report by Pew Research center [3] which polled almost 3000 internet users in its research, provided result that young women are most likely to experience this form of severe targeting from the ages 18-24.With social media being a tool for individuals to express their thoughts and sometimes their emotions. We therefore, need proper internet

surveillance to prevent further escalation of these threat or even flagging down suspicious activities or even profiles.
The aim was to identify these text-based comments by individuals on social media using an approach know as sentimental analysis.
Sentiment analysis can be regarded as techniques used to determine the predisposition of text, which are conveyed in free text form [2]. Subjective data in source materials are selected and gathered with the use of natural language processing, computational linguistics and further text analysis. If is used to uncover an individual's mindset over certain issues or a general polarity through an analyzed text.
This is a proposed technology has sparked a huge interest from academia's for research purposes. With constant research work being carried out in this field of text analysis researchers have discovered more ways and techniques to introduces concept in various fields not only in analytics.
The detection of sexual harassment from social media is an interesting research topic because it serves as a forensic tool in predictive behavioral analysis and patterns from comments online, in this case tweets.

1.1 **RESEARCH QUESTION**

Can the use of supervised machine learning classifiers detect sexual harassment online with the highest accuracy?

This work centers on text mining from twitter, introducing analysis-based sentiments in regards to an individuals' views on twitter and gathering polarity-based score and applying this over various models for accuracy.
The analysis centers on datasets gotten from twitter. Reason for using twitter for the metric analysis is due to its popularity, with users posting general opinion in mini text known as tweets. These tweets are generally restricted to an average of 140 characters and are typically brief which makes it easier for sentimental analysis to be performed.
This would better assist in analyzing the degree of the online victimization from aggressor to the victim. Twitter is used, majorly, due to the following reasons:

- Type of users range from people with different professions, hence there is huge flexibility in the of sources of opinion
- Tweets are straightforward; therefore the authors mindset on a particular subject can be analyzed with ease.

2 Related Work

Online sexual harassment which is a form of cyberbullying is a social issue affecting even adolescents, as the use of the internet grows exponentially. Sexual Harassment, racism, discrimination could cause an individual depression and sometimes even suicides by deteriorating the victim both mentally and physically. It has negative effects not only on the perpetrators but also on others who experience the implications of this violence. It therefore increases crime, mental and physical illness and leads the victims to isolate themselves.
A British survey carried out, recorded that 41% of female consistent Internet users reported being sent unsolicited pornographic contents, harassed, or even stalked on the Internet [6]. These large number of women facing these offensive sex-related experiences on the Internet prompts for further understanding of these misbehavior and others relating to them.

As a response to these cyber threats, a number of national and cross-national child protective initiatives such as The Suicide Prevention Centre Child Focus Initiatives [1] have embarked on projects in the past years to promote online safety. Despite these measures, there is much inappropriate or even hurtful material online.

On average, as reported by new study reports[1], 20 to 40 percent of all teens have been maltreated online. Successful avoidance can be accomplished through the correct identification of potential negative communications. However, there is a requirement for intelligent systems to automatically identify potential risks, given how the internet is overloaded with massive data. This is what encouraged us in introducing sexual harassment tool for detecting these behaviors in different regions and provide a safe environment online. An interesting literature review by Shariff & Gouin which identified that as stated by Herring, 25% of Internet users aged 10-17 who are mostly adolescents were exposed to unwanted pornographic images in the previous year; 8% of the images were violent, in addition to sex and nudity [5]. With these results, this shows that the internet serves as a perfect medium for any age group of digital users to indulge in cyber stalking or harassment.

Although several researches have examined the efficacy of rule-based modelling [8], the prevailing method to cyberSexual harassment detection involves machine learning. Many solutions to machine learning approaches are based on supervised [9, 10–11] or semi-supervised learning [12]. The former requires the creation of a classifier centered on labeled training details, whereas semi-supervised approaches focus on classifiers constructed from a training corpus with a limited set of labels and a large range of unlabeled instances. Semi-supervised approaches are also used to tackle data sparsity, a common problem of research into cyberSexual abuse. Because automated detection basically requires the differentiation between non-sexual and sexual harassment messages, the issue is usually approached as a binary classification function in which the positive class is defined by instances of (textual) cyberSexual abuse, whereas the negative class is devoid of sexual harassment material..

Reynolds and Dinakar investigated the predictive capacity of n-grams (with and without tf-idf weighting), part-of - speech knowledge (e.g. first and second pronouns), and emotion details dependent on (polarity and profanity) lexicons for this role among the first experiments on cyberSexual abuse detection [8–10]. Different apps were used not only to identify coarse-grained cyberSexual harassment but also to identify more fine-grained types of cyberSexual harassment. Notwithstanding their evident simplicity, in recent approaches [12, 14] to its identification, content-based features (i.e., lexical , syntactic, and emotion information) are quite frequently abused. In reality, as noted by [15], more than 41 papers addressed the identification of cyberSexual abuse utilizing content-based apps, suggesting that this sort of knowledge is crucial to the mission.

Social networking is a medium widely used to perform this function. Most recently , researchers investigated the detection of cyberSexual harassment in multi-modal data that specific platforms offer. For starters, Hosseinmardi examined the identification of cyberSexual abuse utilizing multimodal data derived from the Instagram social network[16]. They combined functionality from the posts themselves with user metadata and picture functionality and seek to show that classification efficiency was improved by the incorporation of these. Huang B, Raisi E. CyberSexual abuse was also found in

through data genres like Facebook and Instagram [17]. Role awareness has been taken into consideration by adding bully and victim ratings as attributes, based on the presence of bully-related keywords in their material submitted or obtained.

In regard to the data sets utilized in cyberSexual harassment analysis, it should be found that corpora are mostly constructed by keyword matching (e.g. [18, 20]), which produces a skewed dataset of successful (i.e., sexual assault) incidents. Negative data are also inserted from a historical corpus to establish a certain equilibrium, or data resampling [19] techniques [18, 22] are implemented.

Information is systematically crawled around ASKfm in this study and no keyword analysis was utilized to gather data relating to sexual assault. Alternatively, all cases of the existence of sexual assault is manually annotated. Consequently, our sample includes a rational array of cases of sexual assault. When analyzing the success of automated cyberSexual abuse, we can see the score discrepancies that rely not just on the algorithm and parameter settings applied, but also on some other variables.

These covers the metrics that are used to assess the system (i.e. precision, recall, AUC micro- or macro-averaged F1.), the corpus genre (i.e. Facebook, Twitter, ASKfm, Instagram) and class dispersal based on balanced or unbalanced, the annotation method (i.e. automatic comments or manual comments using crowdsourcing or by experts) and, perhaps the most differentiating element , the theory of cyberSexual harassment that is used.

More categorically, although some mechanisms define sensitive topics[20] or attack comments[19], others recommend a more comprehensive method by collecting various forms of cybervictimization-related comments [20] or by modelling interaction between both the bully and victim which occurs in a cyberSexual harassment incident.

Related research done by [13], [22, 23], where they investigated traces of keywords related to sexual harassment by different roles i.e. the victims, the aggressor and even the passive bystander using Naïve Bayer classifiers only.

However, tweets gathered contained keywords related to bullied, bully and sexual harassment. As a result, their corpus contained detailed reports or testimonials of numerous false positives.

My research topic centers on sexual harassment detection using accuracy score from algorithms such as Random Forest, Xgboost, Support Vector Machine, logistic regression classifiers and Part of Speech tagging for preprocessing.

Large data from twitter will be collected and used for data mining purposes. Keywords which are profane comments related to sexual harassment will be trained and stored. SQLite Database will be used for storing the results.

Twitter API or csv file from Kaggle will be used to gather the corpus to form a data set of three classes

- Positive comment
- Negative comments
- Neutral comments

The main aim of the thesis is to compare various classifiers approaches in our analysis and to determine which algorithm detects sexual harassment with the highest accuracy.

3 Research Methodology

3.1 Statement of the Problem

Given with unlabeled tweet t from user U1 who has the user U2 twitter handle along with a series of past tweets T between U1 and U2, the harassment detection issue aims at automatically detecting whether or not 't' is harassing.

Unlike previous research, this method incorporates past tweets T M amongst specific people. Prior message tweets T play an significant role in the categorization phase because these Tweets may be used to assess the state and intent under which the true tweet was sent.

3.2 Comparison of Machine Learning Classifiers

In the process of testing, a few supervised algorithms were considered, Naive Bayes (NB) which is a very simple algorithm based on probability and counting by condition. A major reason why it was considered was classes of datasets were easy and quick to predict, but a major weakness was if a variable which wasn't observed in training was introduced the entire model will be assigned a zero (0) probability.

Decision Tree algorithm was initially considered due to its strength in performing screening of variables and feature selection. Its major drawback which outweighs advantage as it produces over complex tress which doesn't cause error in the general outcome and this is referred as overfitting.

I. Random Forest: Due to over-fitting being a problem in testing of decision tree algorithm, Random forest was introduced as a fix by averaging several trees bringing about a reduction in the risk of over-fitting. It also builds its model around multiple decision trees and still produces a stable and accurate prediction.

II. XGBoost: The idea behind xgboost is using the computational speed of the system in producing portable and accurate result. It produces its result 5 times faster than other algorithm in testing. Due to its fast computation it can scale well on larger datasets.

III: Logistic Regression: Another algorithm considered because due to its strength in avoiding overfitting and producing an output of probabilistic interpretation. It works great with both linear and non-linear datasets while outputting estimated of instance of variables.

IV. Support Vector Machine (SVM): This algorithm makes use of the kernel mechanism which measures the distance between two views. It then finds a decision limit that maximizes distance between nearest variable of different classes.

4 Design Specification

The entire objective in the research is to build a system based on process learning of massive data and how to categorize it.

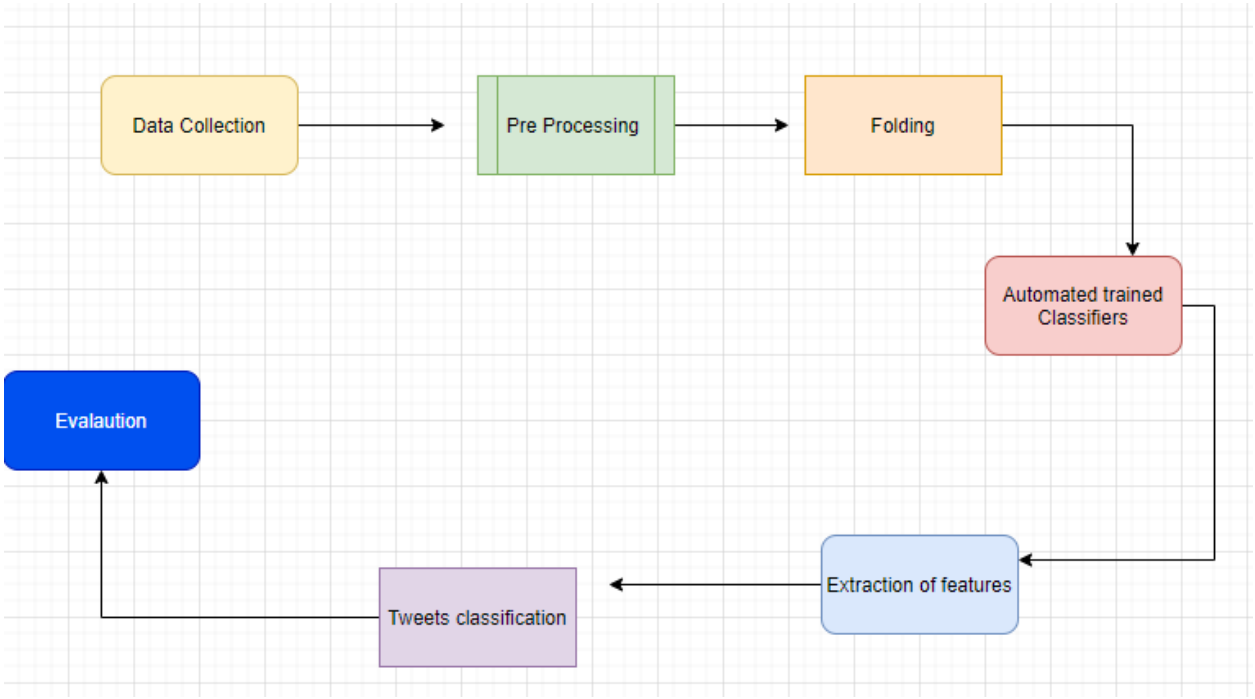


Fig 4.1: Architecture for the proposed implementation

4.1 Data Collection

The first process which involves gathering of relevant data from twitter through the use of api calls. To gain access in developing our and testing our code for our prototype an authorization key was required. With authorization from twitter, tweets were directly collected into our database. Certain keywords were used which are prone to sexual harassment. Keywords such as *dirty whore, ugly, bitch, slut, dirty cunt.* These keywords promote gender bias and are extremely degrading. With the help of the twitter api, corpus was collected and a general dataset of three major classes were generated: negative, positive or neutral sentiments.

4.2 Pre-Processing

For pre-processing, tweets tend to contain spelling mistakes i.e. "*wen r we meetin up*" which are converted in this process to "when are we meeting up". Then uppercase letters are transformed to all lowercase with usernames, URLs and white spaces which aren't needed are all removed.

The use of *stop words* in sentences can be considered useless as they increase the word count without improvement on precision or its recall. These words which can be described as short function word includes: *the, with, is, on* and *for* would be removed in the analysis.

Finally, emojis are deduced using regex on python and theses are coded to classify different expressions. So, expressions such as smiley, blush, eye rolling, angry face or sad face are grouped to corresponding word that relates to the emoticon. Also, emojis which do not relate to a negative or positive classification are either labelled neutral or ignored.

4.3 Folding

Dataset training which certain scenarios involves 80 percent for training and 20 percent for testing. The thesis involves a better practice of 60 percent training, 20 percent each for both testing and cross validation. It makes use of over 10,000 tweets for classifier training and 2000 for polarity testing.

4.4 Automated Trained Classifiers

An automatic process of gathering thousands of tweets while running an algorithm to group each of the words based on their polarity of positivity, negativity or neutrality. A pseudocode of the outlined code is shown below:

```
1.Initialize positiveWordList, negativeWordList
2.positiveScore = negativeScore = 0
3.for each word in tweet
        3.1. if word in positiveWordList:
                3.1.1. positiveScore = positiveScore + 1
        3.2. If word in negativeWordList:
                3.2.2. negativeScore = negativeScore + 1
4.if positiveScore > negativeScore:
        4.1. return 'positive'
5.if negativeScore > positiveScore:
        5.1. return 'negative'
6.if positiveScore = negativeScore:
        6.1. return 'neutral'
```

Fig 4.2 Pseudocode of an automated trained classifiers

4.5 Feature Extraction

- **Repeating Letters**: tweets are stressed to relay emotions i.e. "*I loveeee this"* relates to "*I love this*" and this are converted to the proper sentence during extraction.
- **Punctuation**: punctuation marks such as commas, semi colons or full stops are removed from beginning and end of each word.
- **Stop Words:** Words such as of, with, a, for etc. are useless in labelling the corpus.

4.6 Analysis and Classification

Trained tweets from the database are then classified and run through the four machine learning algorithms to compare these algorithms based on its performance and accuracy.

4.7 Evaluation

Once the implementation of the four supervised machine learning algorithm provides accurate results in the detection on sexual harassment, we analyze how accurate the classifiers are, f1-score and finally the confusion matrix on all the output of the tweets.

F1 score is important as it measures the accuracy and creates a balance between precision and recall doing this.

Finally, confusion matrix which is introduced focuses on the measurement of the performance of the classifiers in predicting correctly or incorrectly.

5 Implementation

For our setup the python package nltk for our classifiers were implemented. Our corpus was labelled to group them into different categories. In our case, in the detection of sexual harassment we need to analyze tweets if they are positive, negative or neutral. These will be passed through a nltk accuracy score. With this score our accuracy calculation can be implemented.

The aim of this thesis is to determine, how accurate these detected negative tweets are determined. With this in mind, a large set of datasets was gathered to train the classifiers to improve the accuracy ratio.

Our computations were generally expressed using NumPy syntax to efficiently executed on the CPU architecture. The NumPy and nltk packages were libraries easily called in our python code.

We have seen that the detection algorithms were used for two things: to classify and predict and that they were divided into supervised and unsupervised algorithms. There are many possible algorithms, we went through 4 of them including logistic regression and random forests to classify an observation then svm and xgboost. We have also seen that the value of an algorithm depends on the associated cost or loss function but that its predictive power depends on several factors related to the quality and volume of data.

The algorithms that we have just seen are used to detect harassment on Twitter, mainly examining the texts of an individual message and also the user information in order to identify harassing tweets. These algorithms mainly use clues such as detecting degrading, insulting or obscene words in a harassing tweet. However, many harassing tweets share this type of trait, these same words end up also appearing as friendly jokes, or as a defense against harassment. These tweets cause false positives and reduce the precision metric of current algorithms. Observation is effective rather than treating individual tweets in isolation to detect harassment, it is more effective to understand the context of a tweet by locating it in relation to the conversation in which it appears.

5.1.1 Tools Used

- Our entire code was implemented on python and run on the jupyter notebook
- Nltk packages for building our python program to work with our human data
- Vader and Text blob were used for testing polarity of negativity or positive tweets
- Scikit machine learning library was introduced for classification based on accuracy metrics on all four algorithms

5.1.2 Corpus

In the bid to analyze harassment, a large group of data (corpus) will be collected for data mining to predict the outcome. In our case, both csv files file Kaggle and twitter api as manual and real time corpus respectively. For our twitter api an authorization key will be needed.

5.1.3 Hashtag

Hashtags are used to find tweets which are related to the subject matter. It is generally used to classify tweets so they are easier to locate. Users use the hash character before a general

topic, caption or post which becomes a trend discussed on twitter. An example is #harrasementatworkplace

5.2 Comparative Analysis

Below we have the tweets and polarity strength and accuracy score of all four algorithms

Tweets	Actual Polarity	Predicted Polarity	score	Random Forest	Xgb Boost	Logistic Regression	Svm
When some women don't know their place and spill rubbish, I say just blow it out your ass	Negative	Negative	72	0.5454	0.8181	0.7676	0.864
I'm sorry don't listen to damn stupid bitches like you	Negative	Negative	65	0.4545	0.5321	0.7272	0.5454
You always remain a black wrench and that's what you always be	Negative	Neutral	27	0.7328	0.6427	0.2382	1.000
She is a minor and it is unacceptable	Positive	Neutral	28	0.4632	0.2311	0.1343	0.3545
Women shouldn't be molested just because of their dressing yes, I said it	Neutral	Positive	23	0.7432	0.8828	0.5410	0.4931

Fig 5.1 Comparative analysis of all four algorithms

6 Evaluation

For the purpose of evaluation, data from our table above were gathered and run through SPSS to get the Precision, recall which was calculated to produce our F1score and plot the AUC Curve. Having a look at the comparative analysis, we have 3 possible outcomes, we expect a confusion matrix to look like the table below.

	Negative	Neutral	Positive
Negative	x1	x2	x3
Neutral	x4	X5	X6
Positive	X7	X8	X9

Fig 6.a: A confusion matrix sample

Assuming that the rows are what was predicted while the columns are the actual

values.

X1 = the number of negative cases predicted as negative

When a case is bad and has been forecast as bad, this implies that the algorithm has correctly categorized it

X2 = the number of neutral cases predicted as negative (Here the algorithm got it wrong)

X3 = the number of positive cases predicted as negative (the algorithm got it wrong too)

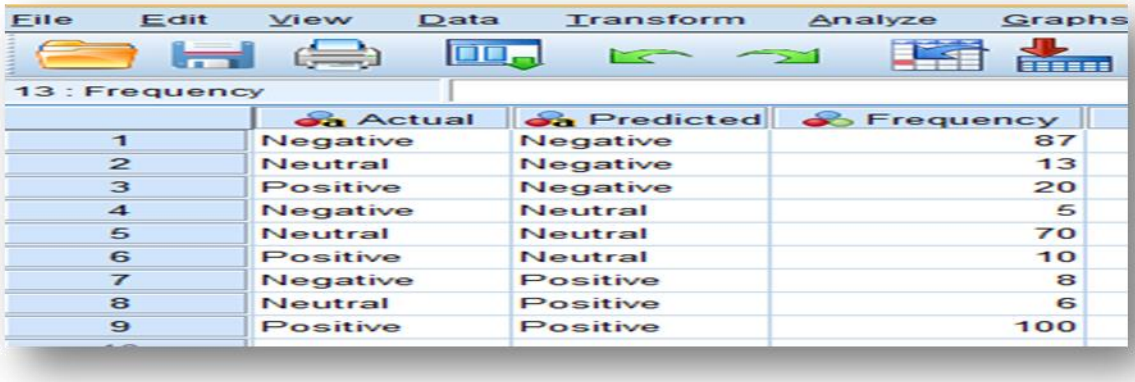
X4= the number of Negative cases predicted as neutral (the algorithm got it wrong too)

X5= the number of Negative cases predicted as neutral (the algorithm got it wright again)

Important facts to take into consideration, the column percentage gives us the sensity and specificity, while the row percentage give recall and precision.

F1 score = 2*(Recall *Precision)/ (Recall + Precision)

For our evaluation output gotten from our program are implemented in SPSS to get both



	Actual	Predicted	Frequency
1	Negative	Negative	87
2	Neutral	Negative	13
3	Positive	Negative	20
4	Negative	Neutral	5
5	Neutral	Neutral	70
6	Positive	Neutral	10
7	Negative	Positive	8
8	Neutral	Positive	6
9	Positive	Positive	100

Fig 6b. Graphing sensitivity against 1- specificity we get the AUC curve

i. Precision

Precision = True positives/all cases predicted positive

We need Row totals for this. The value is 93.8%

1 'Positive' * Actual1 Crosstabulation

			Actual1		
			Positive	Non positive	Total
1 'Positive'	Positive	Count	121	8	129
		% within 1 'Positive'	93.8%	6.2%	100.0%
	Non positive	Count	33	185	218
		% within 1 'Positive'	15.1%	84.9%	100.0%
Total		Count	154	193	347

% within 1 'Positive'	44.4%	55.6%	100.0%

Fig. 6c. Cross Tabulation of Positive and Non-positive

ii. Recall

Recall = True Non positive /all cases predicted Non positive
From the table ***above*** the value is 84.9%

iii. Fscore

F1 Score = 2*(Recall * Precision) / (Recall + Precision)
2*(93.8%*84.9%)/(93.8%+84.9%)= 0.891

AUC

Area under curve gives how good the model is trading off sensitivity to specificity

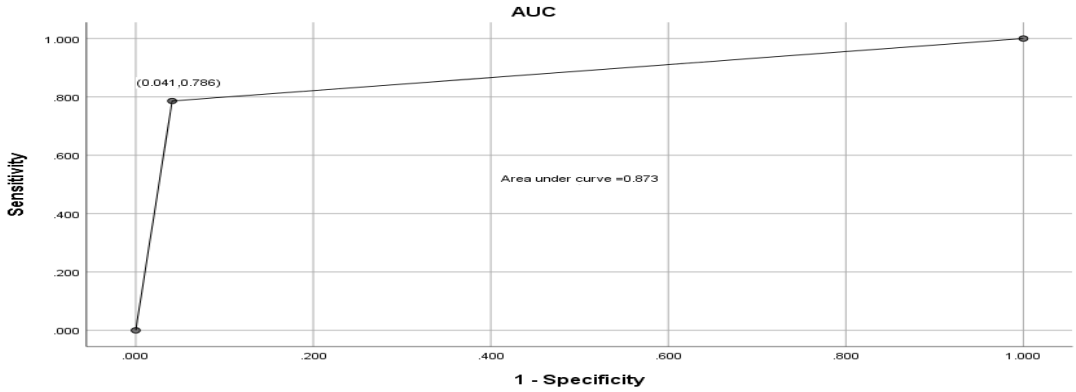


Fig 6d. AUC showing 0.873
If we split this graph into 2 triangles, we can compute the are under the curve

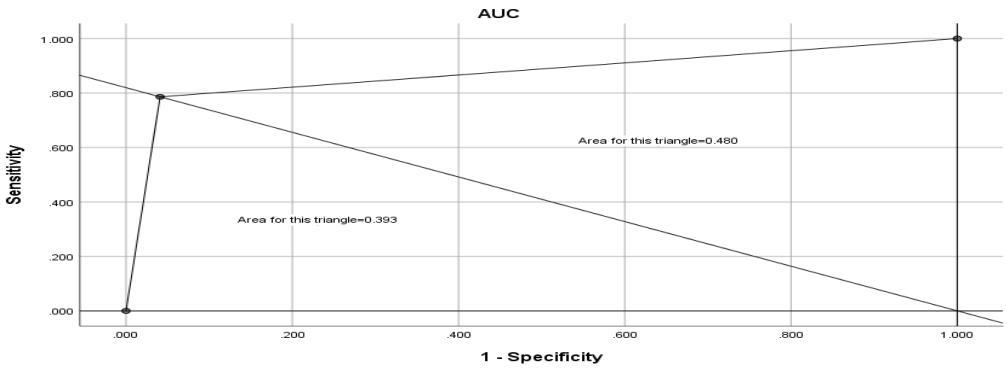


Fig 6e. AUC Sensitivity plot to Specificity with triangle(a)= 0.393 and (b)=0.437

Area of lower triangle is

1/2bh = .5*1*0.786= 0.393
Area of top triangle is
1/2bh = .5*1*(1-0.041) = 0.4375
Total area under curve = 0.873
The general conclusion from our result is the model is good and shows promise.

6.1 Experiment / CaseStudy1

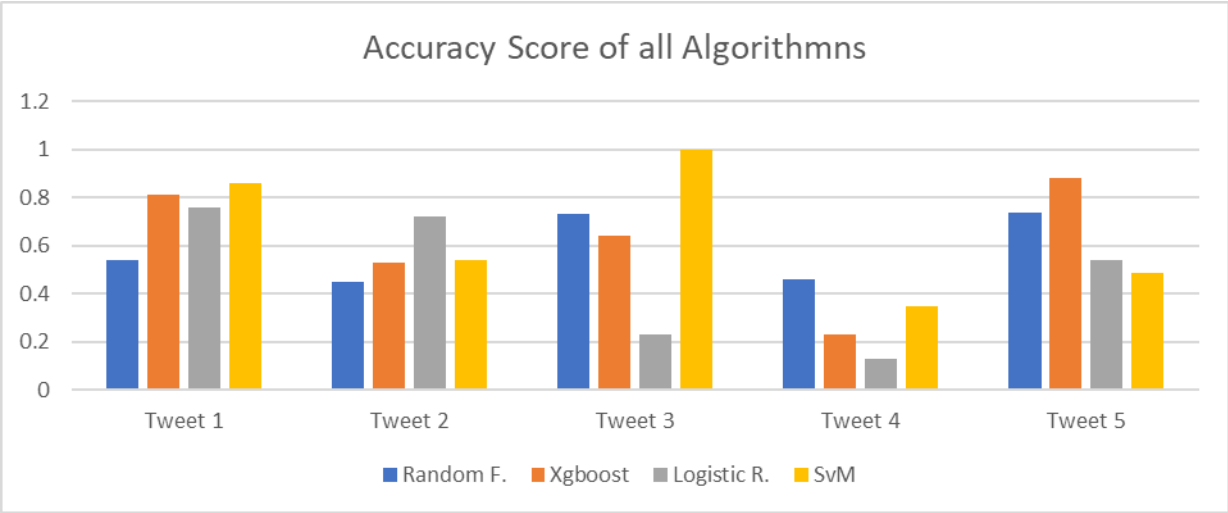


Fig 6f. Accuracy Score of all Algorithms

6.2 Discussion

As we can see from or comparative analysis table and bar chart above, SVM outperformed all other three algorithms in accuracy score. Although over 10,000 tweets were tested, a small sample was used in gathering results of all four algorithm.
From our analysis, Random Forest scored lowest on the accuracy score in predicting negative tweets while scoring higher on tweets considered neutral or positive. Xgboost performed well in detecting negative tweets but outscored still at positive tweets. Logistic regression algorithm performed well in accurately detecting negative tweet with 2 out of 3 negative tweets.
Finally, Support Vector Machine (SVM) showed more promise in detecting negative tweets and ones which aren't in all five tests.
It is vital to understand the effectiveness of classifiers in detecting cyber sexual harassment related tweets accurately. In our measure of the effectiveness of the algorithms sklearn. metrics. accuracy score and NlTK's classify. accuracy was introduced to get the accuracy score on the classifiers.

7 Conclusion and Future Work

Through this research, it's been demonstrated how each machine learning algorithm approach works in its accurate detection of online sexual harassment related tweets. These negative tweets were able to show certain degrees of harassment which is a form of cyber bullying and this was detected successfully with the four approaches- among which, SVM outperformed the rest with Logistic regression coming in second.

In the analysis, the feasibility of accuracy of automatic cyberSexual harassment detection were fully tested with simulations consistently finetuned by repeatedly running various test to produce the best results with the various algorithms. And this can be further perfected in the near future.

The main aim of online sexual harassment detection on social media was to flag a number of possible threats to reduce the effort of manual monitoring of these activities on social network. And this can be achieved by optimizing both precision and recall.

In addition, further classification of tweets based on strength i.e. general insult or even threats can be achieved to provide more insight on what form of harassment is pronounced on social media.

Furthermore, due to the challenges which occur when recognizing certain degrees of online sexual harassment related tweets, further work on the use of advanced functionalities such as syntactic trends and semantic information could be done with Pos tagging implementation.

References

[1] Van Hee C, Lefever E, Verhoeven B, Mennes J, Desmet B, De Pauw G, et al. Automatic detection and prevention of cyberSexual Harrasment. In: Lorenz P, Bourret C, editors. International Conference on Human and Social Analytics, Proceedings. IARIA; 2015. p. 13–18.

[2] Bastiaensens S, Vandebosch H, Poels K, Van Cleemput K, DeSmet A, De Bourdeaudhuij I. CyberSexual Harrasment on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. Computers in Human Behavior. 2014;31:259–271.

[3] "Online Harassment 2017", *Pew Research Center: Internet, Science & Tech*, 2020. [Online]. Available: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/. [Accessed: 03- Apr- 2020]..

[4] Reynolds K, Kontostathis A, Edwards L. Using Machine Learning to Detect CyberSexual Harrasment. In: Proceedings of the 2011 10th International Conference on Machine Learning and Applications and Workshops. ICMLA'11. Washington, DC, USA: IEEE Computer Society; 2017. p. 241–244

[5] Livingstone S, Haddon L, Görzig A, Ólafsson K. Risks and safety on the internet: The perspective of European children. Initial Findings. London: EU Kids Online; 2016

[6] "Sentiment Analysis | Lexalytics", Lexalytics.com, 2020. [Online]. Available: https://www.lexalytics.com/technology/sentiment-analysis. [Accessed: 03- Apr- 2020]

[7] S. Schenk, "Cyber-Sexual Harassment: The Development of the Cyber-Sexual Experiences Questionnaire", ScholarWorks@GVSU, 2020. [Online]. Available: https://scholarworks.gvsu.edu/mcnair/vol12/iss1/8/. [Accessed: 03- Apr- 2020]

[8] Dinakar K, Reichart R, Lieberman H. Modeling the Detection of Textual CyberSexual Harrasment. In: The Social Mobile Web. vol. WS-11-02 of AAAI Workshops. AAAI; 2011. p. 11–17.

[9] Yin D, Davison BD, Xue Z, Hong L, Kontostathis A, Edwards L. Detection of Harassment on Web 2.0. In: Proceedings of the Content Analysis in the Web 2.0 (CAW2.0). Madrid, Spain; 2019

[10] Cortes C, Vapnik V. Support-Vector Networks. Machine Learning.1995;20(3):273-297.

[11] Hosseinmardi H. Survey of Computational Methods in CyberSexual Harrasment Research. In: Proceedings of the First International Workshop on Computational Methods for CyberSafety. CyberSafety'16. New York, NY, USA: ACM; 2016. p. 4–4.

[12] Fekkes M, Pijpers FIM, Fredriks AM, Vogels T, Verloove-Vanhorick SP. Do Bullied Children Get Ill, or Do Ill Children Get Bullied? A Prospective Cohort Study on the Relationship Between Cyberbullying and Health-Related Symptoms. Pediatrics. 2006;117(5):1568–1574. pmid:1665131

[13] O'Moore M, Kirkham C. Self-esteem and its relationship to Sexual Harrasment behaviour. Aggressive Behavior. 2018;27(4):269–283

[14] Zhao R, Zhou A, Mao K. Automatic Detection of CyberSexual Harrasment on Social Networks Based on Sexual Harrasment Features. In: Proceedings of the 17th International Conference on Distributed Computing and Networking. No. 43 in ICDCN'16. New York, NY, USA: ACM; 2016. p. 43:1–43:6.

[15] Raisi E, Huang B. CyberSexual Harrasment Detection with Weakly Supervised Machine Learning. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017. ASONAM'17. New York, NY, USA: ACM; 2017. p. 409–416

[16] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer PW. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research (JAIR). 2002; 16:321–357.

[17] Slonje R, Smith PK. CyberSexual Harrasment: Another main type of Sexual Harrasment? Scandinavian Journal of Psychology. 2008;49(2):147–154. pmid:18352984

[18] Nandhini B Sri, Sheeba JI. Online Social Network Sexual Harrasment Detection Using Intelligence Techniques. Procedia Computer Science. 2015;45:485–492

[19] researchgate.net,2020.[Online].Available: https://www.researchgate.net/publication/228799343_Cyber_bullying_Clarifying_Legal_Boundaries_for_School_Supervision_in_Cyberspace. [Accessed: 03- Apr- 2020]

[20] Chang CC, Lin CJ. LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011;2(3):27:1–27:27.

[21] Dadvar M. Experts and machines united against cyberSexual Harrasment [PhD thesis]. University of Twente; 2018

[22] Nahar V, Al-Maskari S, Li X, Pang C. Semi-supervised Learning for CyberSexual Harrasment Detection in Social Networks. In: ADC.Databases Theory and Applications. Springer International Publishing; 2019. p. 160–171

[23] O'Sullivan PB, Flanagin AJ. Reconceptualizing 'flaming' and other problematic messages. New Media & Society. 2003;5(1):69–94.