2023

Online Sexual Predator Detection

Muhammad Khalid
University of Windsor

Online Sexual Predator detection

By

Muhammad Khalid

A Thesis
Submitted to the Faculty of Graduate Studies
through the School of Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science
at the University of Windsor

Windsor, Ontario, Canada

2023

Online sexual predator detection

by

Muhammad Moeed Khalid

APPROVED BY:

_____

J. Pathak
Odette School of Business

_____

D. Alhadidi
School of Computer Science

_____

A. Ngom, Co-Advisor
School of Computer Science

_____

H. Fani, Co-Advisor
School of Computer Science

January 18, 2023

Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Online sexual abuse is a concerning yet severely overlooked vice of modern society. With more children being on the Internet and with the ever-increasing advent of web-applications such as online chatrooms and multiplayer games, preying on vulnerable users has become more accessible for predators. In recent years, there has been work on detecting online sexual predators using Machine Learning and deep learning techniques. Such work has trained on severely imbalanced datasets, and imbalance is handled via manual trimming of over-represented labels. In this work, we propose an approach that first tackles the problem of imbalance and then improves the effectiveness of the underlying classifiers. Our evaluation of the proposed sampling approach on PAN benchmark dataset shows performance improvements on several classification metrics, compared to prior methods that otherwise require hands-crafted sampling of the data.

Keywords: Natural Language processing, Machine Learning, Deep Learning, Data Imbalance, Classification

Dedication

I dedicate this thesis to my family, my mother, my father, and my sister, who have supported me through every step of my life and every decision I have ever made. I could not have done this without you.

And to my friends, who have played a vital part in me becoming the person I am today.

Acknowledgements

I would like to thank my supervisors Dr. Alioune Ngom and Dr. Hossein Fani, without whose guidance none of this would have been possible, and who were extremely patient and compassionate with me through a very difficult phase of my life. Thank you for pushing me to always do my best.

I would also like to thank my internal reader Dr. Dima Alhadidi and my external reader Dr. Jagdish Pathak for their valuable insight and suggestions on my research.

I would like to thank my mother, without her encouragement and love I would not be at the position in my life where I am today.

Table of contents

List of tables

List of figures

Chapter 1

Introduction

1.1 What is a predator?

To understand what a predator is, it is important to first understand the definition of a paedophile. Any adult that feels the urge to have sexual advances towards a minor regardless of the gender or demographic, is known as a paedophile [1]. These feelings can be towards children that those paedophiles know or might happen because of certain stimuli. This is the closest definition we have of a paedophile as there has been very less research in this regard and empirical analysis yields this to be the best description [2].

Now that we have established what a paedophile is we need to define what a predator is. The main difference between a paedophile and a predator is that while paedophiles have sexual feelings towards minors, they do not necessarily act upon it. Whereas predators are paedophiles who give in to the urge and actively prey on vulnerable children [3].

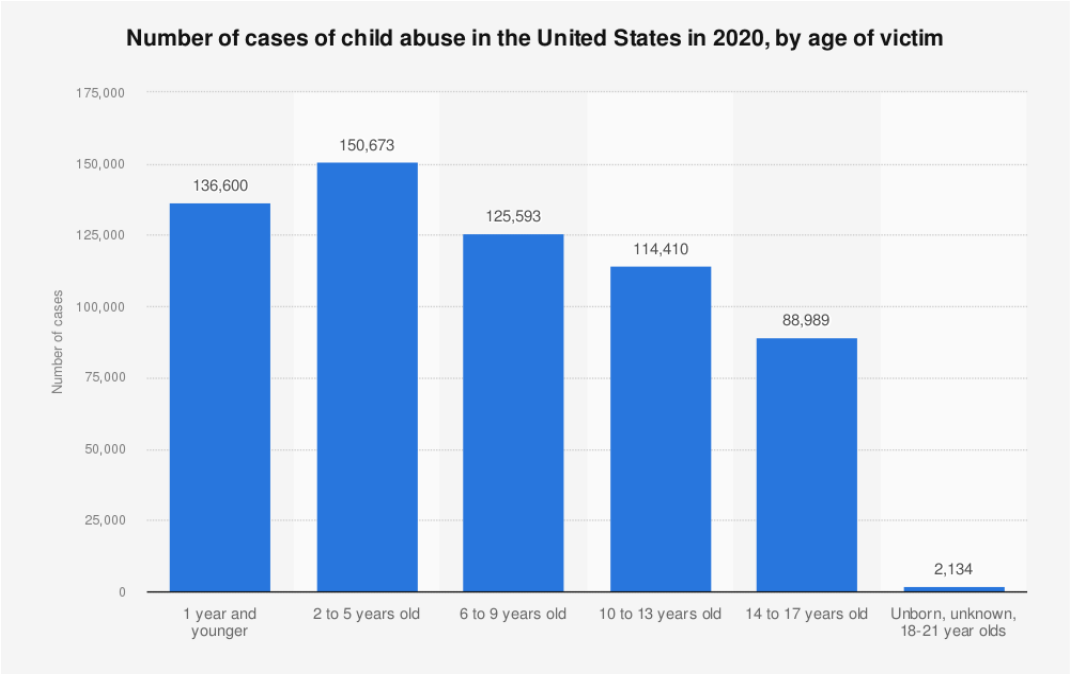

Figure 1.1 Child abuse statistics in the United States

On the surface it might seem like a problem that is not that prevalent as it hasn't been talked about much in mainstream media, but figure 1.1 depicts the alarming child abuse statistics in the United States only.

1.2 How Predators prey?

Predators have a very measured approach towards targeting a child. They do so via a process called grooming. The definition of grooming is not exactly agreed upon but lets take a look at some different interpretations of the word and try deduce what the word means.

1.2.1 Grooming

O' Connell defines the act of grooming as:

> *"A course of conduct enacted by a suspected pedophile, which would give a reasonable person cause for concern that any meeting with a child arising from the conduct would be for unlawful purposes"* [4].

Another definition given by Hawitt is:

"Grooming ... is the steps taken by pedophiles to "entrap" their victims and is in some ways analogous to adult courtship [5]*"*

In 2001 Gillespie refined the definition to say:

"*The process by which a child is befriended by a would-be abuser in an attempt to gain the child's confidence and trust, enabling them to get the child to acquiesce to abusive activity. It is frequently a pre-requisite for an abuser to gain access to a child.*" [6]

From the above definitions we can draw a parallel and generalize to say that grooming is the pre-requisite to the actual heinous activity that predator would do to their would be prey. Our research is based to catch the predator in the grooming stage of the process where the child has yet to bear physical harm from the hands of the predator.

1.2.2 Stages of Grooming

Grroming is not a straightforward process. So, to better understand what grooming is the whole process is divided into several portions called the stages of grooming [7]. Even

then the extent as to how much the child is entrapped in each of these phases has not been agreed upon [8].

The following are the 6 stages of grooming [9]:

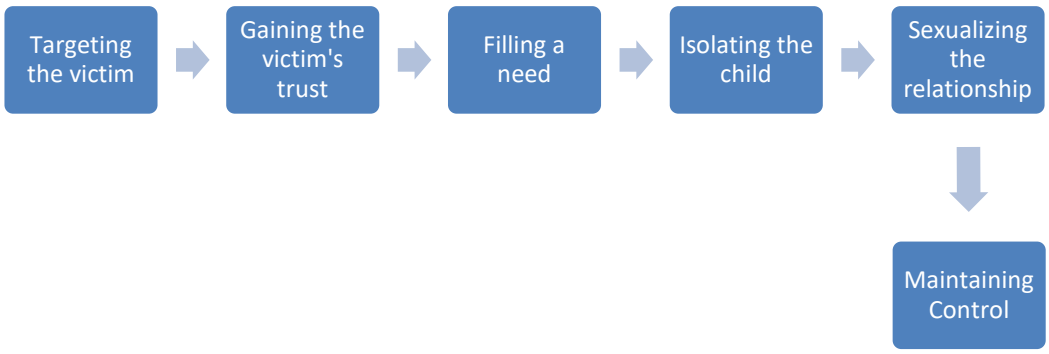

Figure 1.2 Six stages of grooming

Targeting a victim: The predator would look for emotionally vulnerable children. In most cases this will be children that are facing problems in their real life and are looking to fill a void.

Gaining the victim's trust: In this step the predator would try and get familiar with the child, get as personal as they can with them while making them feel safe. They do so to gain the victims trust so the victim does not feel any hesitation talking to them going forward.

Filling a need: Once the predator has gained the child's trust, they will then try to be as affectionate to them as they can in the manner that the victim feels like they are the figure that is going to fill the void in their lives.

Isolating the child: The predator would then isolate the victim by giving them even more attention and creating scenarios in which the child feels like they are forming a bond with the predator on an individual and an isolated basis.

Sexualizing the relationship: Once the predator has control over the child, they would then start sexualizing the relationship. This can be done by asking for inappropriate photographs or talking to them about any situation in which the victim is naked such as their swimming trips etc. They do so to advance the relationship in a more sexual manner.

Maintaining control: Once the relationship has gone sexual the predator would then do everything to keep the child under their control. That means they would employ any tactic at hand, whether it be scaring them or manipulating them, to keep them under control so that the predator can keep them emotionally entrapped.

1.3 Key Definitions

Following are some of the key definitions and terminologies that we will use throughout this discussion.

Predator: The person in the conversation who is trying to prey on the children online.

Prey: The minor who is being preyed upon in the conversation.

Grooming: The exchange of emotional/sexual messages between the predator and the victim.

Conversation: A collection of back-and-forth messages between two or more people.

Predatory Conversation: A conversation that has at least one predator involved in it. So, any conversations in which a predator plays any part in is considered a predatory conversation.

Message: Every individual message in a conversation.

Predatory Message: Every message that has been sent by a predator. This is the message that we target on classifying.

1.4 Thesis Motivation

With the advent of more technology, everything in our life is becoming more and more digital. Children have access to technology way before they are of legal age [10] and have very little cognitive development. One of the more alarming problems that is faced in this regard is children interacting with predators on- line in sexual grooming conversations. Deep web used to be a hub for illegal activities [11] that include but are not limited to human trafficking, organ smuggling, child pornography etc. [12]. But in recent times mainstream digital platforms such as online video games [13] and chat rooms [14] are some of the more common places where children are present, more often than not, and are easy prey for online sexual predators such as those who are diagnosed with pedophilia. The ways that these adults engage in this act differ in a lot of ways such as using familial power or localized pressure [8]. In many of these instances, the offenders try and mix in explicit remarks in the conversation to get a sense of how they are going to proceed talking to the victim. We can exploit this attribute to catch such offenders. Natural Language Processing (NLP) and Machine learning are two of the most sought out methods in this regard. According to a case study [15], around 60-80% of female high school students have to face online sexual grooming incidents. Getting a good grasp of identifying conversations of grooming nature can result in these predators getting caught and can in turn save countless children from getting scarred for life.

Kids between the ages of 12 and 15 are very prone to being groomed by online sexual predators. According to the report from the FBI 89 % of all sorts of sexual advances towards children happen online as shown in Figure 1.4.

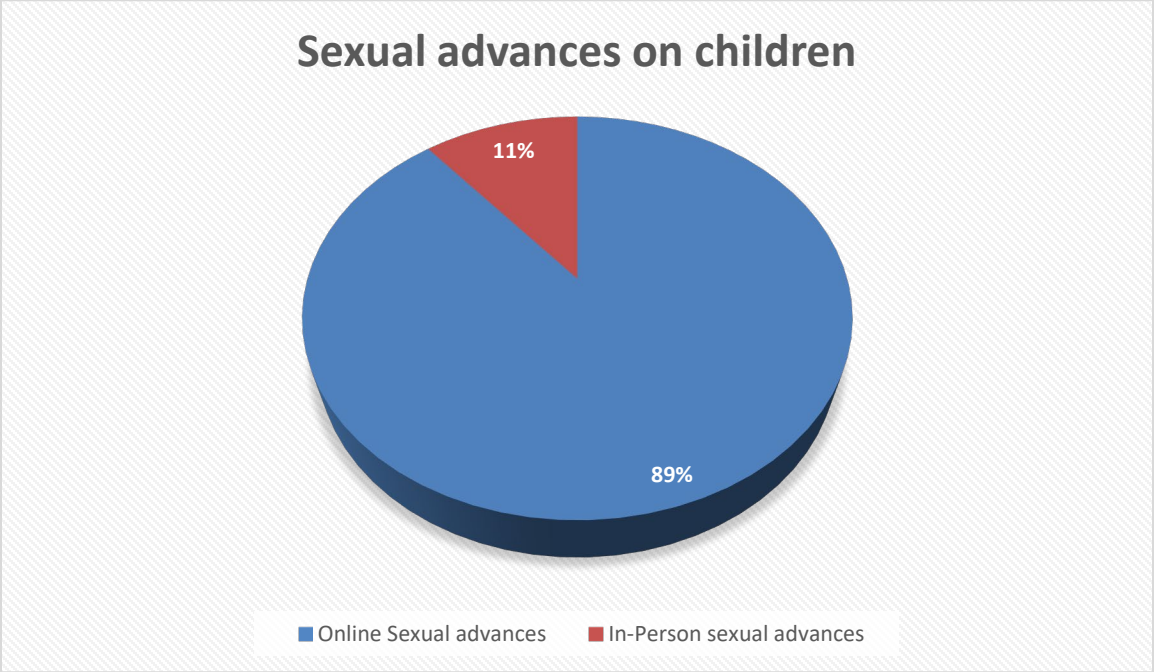

Figure 1.3 Sexual advances on children

1.5 Motivation Example

Below is a snapshot from a conversation between a predator and an underage victim. This is the tip of the iceberg of the actual conversation that happens between them in the more than 300 message long thread.

```
<conversation id="conv_snippet_end">
  <message line="25">
    <author>victim</author>
    <time>10:21</time>
    <text>ima goin swiming for a lil im hot</text>
  </message>
  <message line="26">
    <author>predator</author>
    <time>10:27</time>
    <text>yeah you are..... ohhhhhh.. the other "hot"</text>
  </message>
  <message line="27">
    <author>predator</author>
    <time>10:27</time>
    <text>gotcha ;)</text>
  </message>
  <message line="28">
    <author>predator</author>
    <time>10:27</time>
    <text>bye hun</text>
  </message>
</conversation>
```

Figure 1.4 Sample Predatory Conversation

1.6 Thesis Contribution

In this thesis, we have proposed a method to tackle the problem of data imbalance in a dataset of sexually predatory conversations and detect sexually predatory messages in a more automated manner.

By this automation of the process of detection of online predators, we want to speed up the process of catching predators in online conversations. As a result of this, even if our method can make the process 0.1 % more efficient, this means for every 1000 children that were going to fall prey to a predator we can save one of them, which means we can

save one life from getting scarred for life and that is the most important part of this research.

1.7 Problem Definition

Given a corpus of conversations C with every individual conversation c, we wanted to find grooming messages m such as $m \in c$.

1.8 Thesis Organization

This thesis is organized in the following manner:

- **Chapter 2**: We discuss related work that is already done in the field of online sexual predators' detection
- **Chapter 3**: We explain our proposed method and our dataset
- **Chapter 4**: We discuss all the experiments that were performed and their results
- **Chapter 5**: We draw conclusions from our research and talk about possible future work.

Chapter 2

Literature Review

In this section we will be taking about different types of methodologies that have been used to tackle online sexual grooming methodologies, we will touch on the datasets that they have used and the algorithm that they have proposed. Approaches to identify predators include labelling a conversation [16] as well as labelling each sentence of the conversation [17].

2.1 Shallow Learning Methods:

2.1.1 Detecting Child Grooming Behaviour Patterns on social media

Paedophilia is being focused in this paper by Cano et al. [18]. They used the dataset provided by the perverted justice foundation [19]. This dataset contains text from chatrooms where adults pose as underage children to try and find predators so that they can be bought to justice. As of January 2021, this foundation has been responsible for convicting 622 predators.

Conversations that are of a grooming nature or involve sexually explicit content are then added to the dataset, so it keeps on growing. This paper uses NLP techniques on different features such as lexical, pshycolingual etc. to get the best performance measures out of the dataset.

2.1.2 Toward spotting the pedophile telling victim from predator in text chats

Working on the same perverted justice foundation dataset Pendar et. al [20] used machine learning algorithms to approach the problem. He used SVM and k-NN algorithms to identify chats that were of paedophilic or grooming nature. The measure of performance in this paper was micro and macro average precision and recall. Their best results for the k-NN were achieved using 3-gram with N=3. The website of the dataset gives the average public a chance to vote how slimy they think the text of a chat is so this paper also takes into consideration the "sliminess" of the predator, which means how sexual the content of the conversation is.

2.1.3 Learning to identify internet sexual predation

Mcghee et al. [21] used a decision tree to classify whether or not the conversation under discussion is of a grooming nature. They have specified words that if those words are present in a conversation, then that conversion is going to be deemed as a grooming conversation. The accuracy for this decision tree on this dataset was relatively low as compared to other methodologies.

2.1.4 Identifying online sexual predators by SVM classification with lexical and behavioral features

Morris [17] also worked on the Pan-12 dataset by implementing an SVM and using precision and recall as accuracy measures. What Morris has done in this paper is create numeric vectors of data and then pass them through a multi layer perception. Of all the papers being discussed this is the only one who uses this technique to preprocess the data. Another thing being done in this paper is it is assigning a label to each of the sentences. The linear SVM tags each line with the level of "predatoriness" of the sentence.

2.1.5 Conversation Level Constraints on Pedophile Detection in Chat Rooms

Peersman et al. [22] participated in the PAN-2012 competition and gave forward their solution which was a multi-step process. Firstly, they tagged the whole conversation that whether or not this conversation is of grooming nature. They did so by applying an SVM. Then they moved on to the user level classification. In this step they put tags on the users that were present in the conversations that were of a grooming nature. They labelled them as either predator or a victim.

2.1.6 Characterizing Pedophile Conversations on the Internet using Online Grooming

Gupta et al. [23] also uses the Perverted Justice dataset to detect predators. This paper takes a different approach towards tackling the issue. They manually label a select number of conversations from the dataset. They don’t just put labels to the whole conversation. They label it in the stages of grooming, from targeting a victim to gaining control [24]. Every portion of the text is labelled according to what phase of grooming it belongs to. After that not only do they detect what phase of the grooming conversation is under conversation but also tell if the current conversation has a phase of transition

between two stages. This correlation between stages is done via LIWC which is a widely used word counting program.

2.1.7 Predatory Conversation Detection

Borj et al. [25] used the PAN2012 dataset and proposed a study on the detection of predatory conversation and used a variety of methods to detect cyber-criminal activity, including linear SVM, SVC, Naive Bayes, and Random Forest. Researchers also looked at various aspects of online conversation, such as psycho-linguistic patterns, when analyzing the various types of features of grooming data. As per the findings, Predatory conversations were recognized with 98 percent accuracy by linear SVM and NB because of their experiments, with linear SVM having a better F-score of 0.84 for predatory talk detection. In addition, they achieved the best results when using multinomial naive Bayes on 1-gram features and linear SVM on 1-gram features when stop words were not removed.

2.1.8 Detection of Cyber Grooming in Online Conversation

Bours et al. [26] has recently used a new approach for the classification of the documents. They have used the PAN-12 dataset and tried to compare the Bow and TF-IDF feature sets on different classifiers. They have compared different machine learning algorithms; namely: Logistic Regression, Naive Bayes, Ridge, SVM and Neural Networks. They have used precision and recall as their accuracy measures. One of the unique things that they have done in their paper is that they have tried to trim the documents from the bottom, the logic being that they want to find the minimum possible range from the first message in which they can get the best classification measures for their dataset. They found out that in the case of feature sets, TDF-IF outperforms BoW in all scenarios. As for the machine learning algorithms Ridge and Naive Bayes give the most accurate results.

2.1.9 A Simple Classifier for Detecting Online Child Grooming Conversation

Gunawan et. Al [27] used two diverse types of conversations in this study: actual online child grooming conversations and non-grooming conversations. The first type of conversation was chosen at random from. They have used SVM and KNN to find the potential for online child grooming conversations. On the basis of the quantity of existing

grooming conversation characteristics, the study also suggests a classification method with a low computational cost. In addition, 45 non-grooming texts and 105 grooming texts from 150 conversation texts were used to evaluate each suggested method. Finally, the analysis shows that grooming conversations contain 17 grooming traits

2.1.10 A Two-step Approach for Effective Detection of Misbehaving Users in Chats

The Language Technologies Lab at INAOE and the Language and Reasoning Group at UAM collaborated on a solution for the PAN 2012 Sexual Predators Identification task from the paper [28]. The proposed technique addresses the challenge of spotting sexual predators in a group of questionable chats. Their goal was to show that it is possible to train a classifier to learn the specific terms that turn a chat conversation into an instance of online child exploitation, as well as to learn the predators' behavioral patterns during a chat conversation. It is allowing us to accurately distinguish victims from predators. The methods they have used: are Neural Networks (NN), Support Vector Machines (SVM) from the CLOP toolbox, and two-fold cross-validation to estimate the performance. The SVM accuracy is 0.97 and the neural network accuracy is 0.99. Additionally, the researchers' involvement in the PAN 2012 forum demonstrated that the suggested methodology is capable of producing excellent results in a realistic scenario, as evidenced by the F-measure (= 0.5) of 0.8936, which was the highest-ranked outcome among all of the participants.

2.1.11 Sexual-predator Detection System based on Social Behavior Biometric (SSB) Features

On the PAN 2012 corpus, the tests are performed. The goal of social biometrics is to figure out how a user interacts and communicates on social media platforms. Researchers used vocabulary and emotional behavior analysis to decide whether a user is benign or predatory, which helped them solve the problem of online sexual predators in the paper [29]. In this paper, the training data has been split into two sets with an 80:20 split between the training and validation sets. They used Decision Tree (DT), SVM, and Random Forest (RF), and the SBB-based approach had the best accuracy with 99.86, 99.51, and 99.88 percent, respectively, because of their experiments. In addition, as per the findings, the results of these test sets have demonstrated that the system's

performance has been significantly improved. By obtaining F1, F2, and F0.5values of 0.95, 0.94, and 0.96 respectively, the suggested technique has surpassed the best current techniques when compared.

2.1.12 Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques

The data for this study came from "maps.safe city," which is a tracking system for online crimes like cyberbullying, domestic abuse, molestation, and sexual assault as mentioned in the paper [30]. Basically, cyberbullying is regarded as a dangerous human activity that takes place online and easily harms innocent users, authorities, or other targets. Thus, the primary goal of this research was to propose a technique for utilizing machine learning algorithms to enhance the classification of various types of malicious human activities as well as to create detection systems. To test the proposed model, researchers used the same training and testing datasets with eight different classifiers, including Random Forest, Multinomial Naïve Bayes, SVS, Linear SVC, SGD, Bernoulli NB, Decision Tree, and K Neighbors. According to the research, tests revealed that combining Term Frequency Inverse Document Frequency (TF-IDF) with machine learning led to an accuracy rate of 81 percent.

2.2 Deep Learning Methods:

2.2.1 Classification of Predators using Convolutional Neural networks

Ebrahimi et al. [16] is the only paper to propose a Convolutional Neural Network (CNN) to tackle the problem of detecting predators from online conversations. They have used the PAN-12 dataset which is an extension of the Perverted Justice dataset. They used F1 as a performance measure and got its value closes to 80 %. One of the things that they did in creating their CNN that stood out was they did not preprocess their data in any way shape or form. Also, they have used only one layer in the CNN which is an interesting approach as they claim that adding more layers on these predatory datasets (which are relatively small) results in the data overfitting.

Chapter 3

Proposed Method

In this chapter, we will discuss in detail, our dataset, data preprocessing, data imbalance handling technique as well as the machine learning and the deep learning models used in our thesis.

3.1 Dataset

The dataset that we use in our experiments is the PAN-2012 predator identification dataset [31]. This is one of the most extensively used dataset in the research area of online sexual predator detection as it is the only one of this magnitude available in English language to date. Almost all the predatory messages in this dataset are provided by the perverted justice foundation which has one of the biggest logs of chats between predators and victims, whereas the rest of them are gathered from Omegle chat rooms which is a website where strangers can chat while having the option to turn on video chat as well.

We will notice a severe imbalance between the data, this is because the authors of the dataset wanted there to be conversations from all fields of life, so there are a lot of non-predatory conversations in the dataset as compared to the few predatory that they got from a perverted justice foundation and Omegle.

The dataset contains about 3 million messages, that are gathered from different chat rooms with an intent to identify predatory messages and conversations.

Some of the stats about the dataset are given in Table 1. The dataset is severely imbalanced in favor of non-predatory conversations. Only 2.3 % of the conversations are predatory conversations and only 0.12 % of the users are predators. This is an extreme skew towards non predatory data.

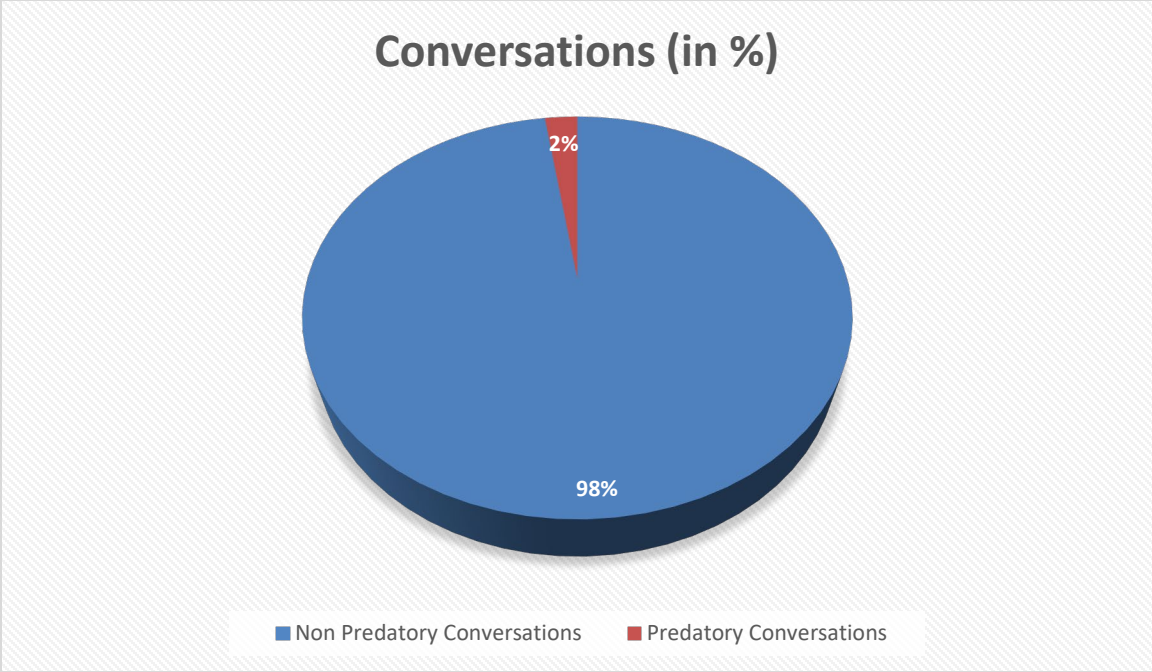

Figure 3.1 Imbalance between predatory and non-predatory conversations

Some other facts that we discovered about the dataset include that there are no predators in non-binary conversations, meaning that any conversation that involves more than one participant does not have a sexual predator in it and there is no more than one predator in a conversation.

Total	Training Data	Test Data
Conversations	66,927	155,128
Predatory conversations	2,016	3,737
Messages	903,607	2,058,781
Users	97,689	218,702
Predators	142	254
Average messages	13.5	13.27

Table 3.1 Dataset Statistics

3.2 Proposed method

The approach we propose involves the following steps. Training a Word2Vec and getting the word vectors for each sentence, handling the data imbalance (Synthetic Minority oversampling technique [32]), feeding the data to different Machine Learning and Deep Learning algorithms and assess their performance using Area under the curve. Let's discuss in detail all these three things.

3.2.1 Word2Vec

To understand word2vec let us first talk about word embeddings and why we need them. Word embeddings simply put, is the method of building vector representation of words [33]. There are several ways by which we can create vector representations of words or documents. Some of the popular ones include one hot encoding [34] which involves giving every single word, alphabet, or pixel its own unique value and repeating that whenever that datapoint occurs again. It is one of the most commonly used techniques to encode small documents. But with a dataset of almost a million sentences, this unfortunately did not work for us.

We tried using doc2Vec [35] in which process each document (which in our case is a sentence) is tagged as a vector and that is fed to the neural network. Using doc2vec was giving us sub-par results so we trained our own word2vec [36] and used that which gave us our best results.

Word2vec is a word embedding technique in which every single word in our corpus, including or excluding stop-words, depending on how you want to train your model is given a vector representation. This representation is based on the word and its semantic relationship with other words that in the documents it occurred in. A good testament to whether or not a word2vec model accurately represents the corpus it was trained on is to find words that are thematically relevant to the corpus and trying to find similar words, if the similar words are semantically similar and comparable to the input word, that means that the word2vec was well trained and an accurate representation of the corpus.

The word2vec can be trained either using skip gram or Continuous Bag of Words (CBOW) representation [37]. We used CBOW to train our w2v model. CBOW places the word in the center and uses the past and future words to try and draw a semantic relationship and infer the word. It does so by projecting all those words using a weight matrix and then calculating the softmax for that distribution.

3.2.2 Synthetic Minority oversampling technique (SMOTE)

Machine learning as well as deep learning algorithms underperform when the data is too tilted towards one of the classes. This data imbalance, specially in datasets that have overlapping data results in models severely underperforming in most cases.

Figure 3.2 illustrates the imbalance that our dataset had, for about 1.5 million negative samples, we had only about 60 thousand positive ones. This magnitude of data imbalance results in the training of garbage models.

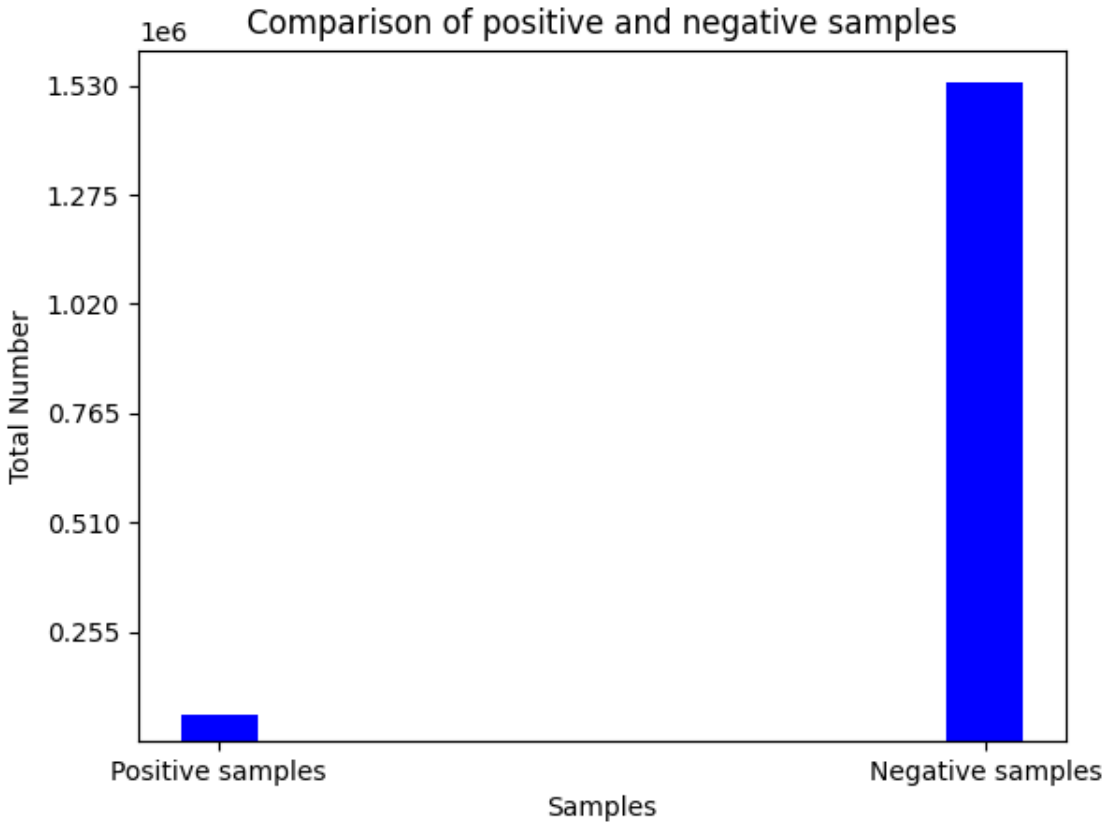

Figure 3.2 Comparison of positive and negative samples in unbalanced data

There are several techniques that are used throughout the machine learning world to handle class imbalance, these techniques are branched out of methods like oversampling, resampling, ensemble different datasets etc.

One such techniques that is widely used is synthetic minority oversampling technique or SMOTE [32]. This approach was proposed keeping in mind that just replacement oversampling of the data does not always result in an improved performance.

SMOTE tries to find more synthetic way getting more datapoints of the minority class. SMOTE starts out by at a datapoint of the minority class, it then finds out its K nearest neighbours typically the value of k is either taken as 3 or 5, then there is a line drawn between our datapoints and its k nearest neighbors, SMOTE then creates datapoints along that line, this results in a more synthetic way of creating samples for the minority class rather than replacement oversampling.

For our dataset we applied SMOTE to balance out the datasets and as can bee seen in figure 3.3, we got 1.5 million endpoints of both the positive as well as the negative samples and we continued our experiments afterwards.

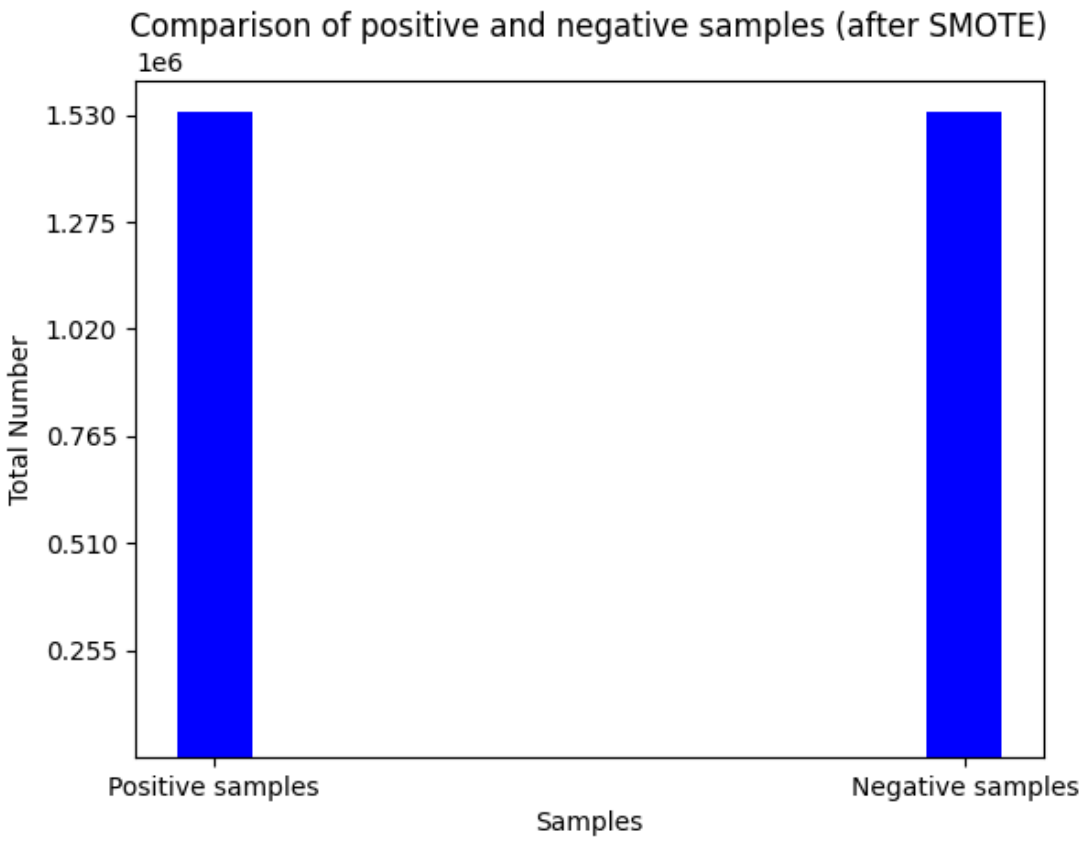

Figure 3.3 Comparison of positive and negative samples after handling data imbalance

3.2.3 Baselines

We will be using several Machine Learning algorithms and compare their performance on our dataset as well as compare it to some deep learning approaches we will be discussing in this thesis. Namely we will be using Gaussian Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Random Forest, and K-Nearest Neighbors. Our dataset has a severe imbalance as it can bee seen in the dataset section, so the metric that we will be using to evaluate our dataset is going to be the Area under the curve. Scikit-learn's implementation of these algorithms is being used in all our experiments. Models we have used are:

Gaussian Naive Bayes:

An extension of the Naive Bayes algorithm that uses Gaussian normal distribution and works on continuous data as well [38]. Most of the classification tasks performed by Naïve Bayes algorithm include spam filtering, document classification etc.

The classifier is named after Thomas Bayes who put forward the Bayes theorem. Bayes theorem simply put, is the probability that an even would occur given that the probability of another event that has already occurred

In Naïve Bayes algorithm, it is called a Naïve algorithm because the events of one probability do not affect the events of the other one. Similarly Naïve Bayes algorithm works on the heavy assumption that two features that are used for the purpose of classification are completely independent from each other such that one feature being present or not does not affect the other features in any way shape or form.

Linear Discriminant Analysis:

A method used to classify two or more classes using a linear separation technique [39]. It is an extension of Fisher's linear discriminant which is used to separate two or more classes. Linear discriminant analysis is a really good baseline method and can be really fast and efficient but fails when there is a lot of overlap between the datapoints of different classes. It works best on data that is linearly separable.

Logistic Regression:

Logistic regression [40] is one of the most used classification algorithms in the field of Machine Learning. Although it is able to perform regression tasks, but it is widely used for classification purposes. Logistic regression fits a line most commonly which is curve fitted. This gives us continuous values as a result which is then translated into a binary output by the means of logistic function.

Random Forest:
Random Forest is a popular ensemble Machine Learning technique used for Classification as well as Regression. It is an ensemble classifier that makes predictions

using a variety of decision trees [41]. It fits various decision tree classifiers to different dataset subsamples. Each tree in the forest was created using the best random subset of features. One of the major benefits of Random Forest is that it does not faulter against extremely large datasets and can perform really well in scenarios where data is even missing.

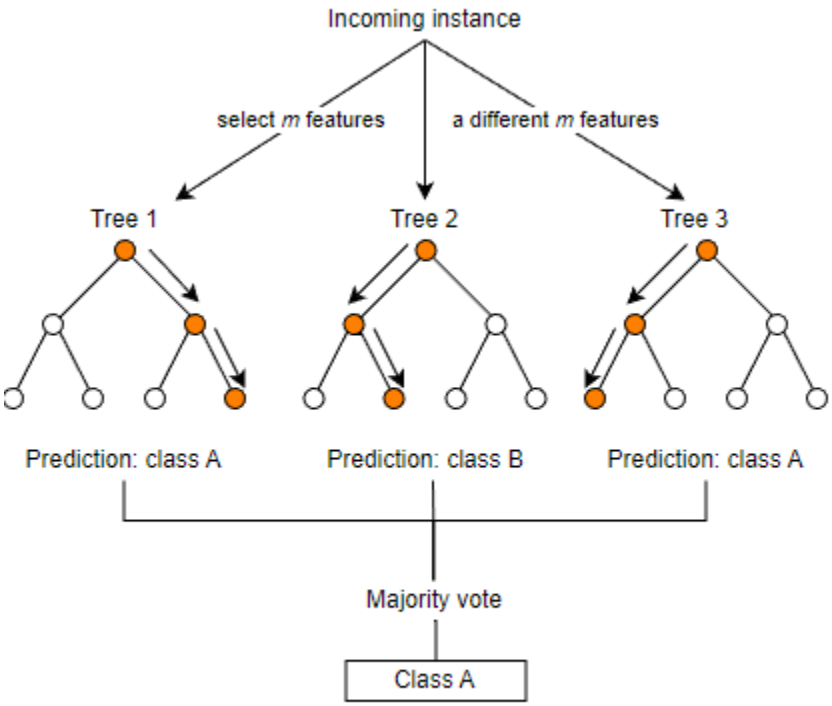

Figure 3.4 Depiction of Random Forest classifier

3.2.4 Deep Learning Models

Deep learning is a branch of Machine learning in which we try and build models as close to human how a human knowledgebase works as we can [42]. We do so by creating models that have multiple layers so that we can then define and refine according to our needs. These methods are some of the most computationally intensive methods out there and in turn usually yield better results.

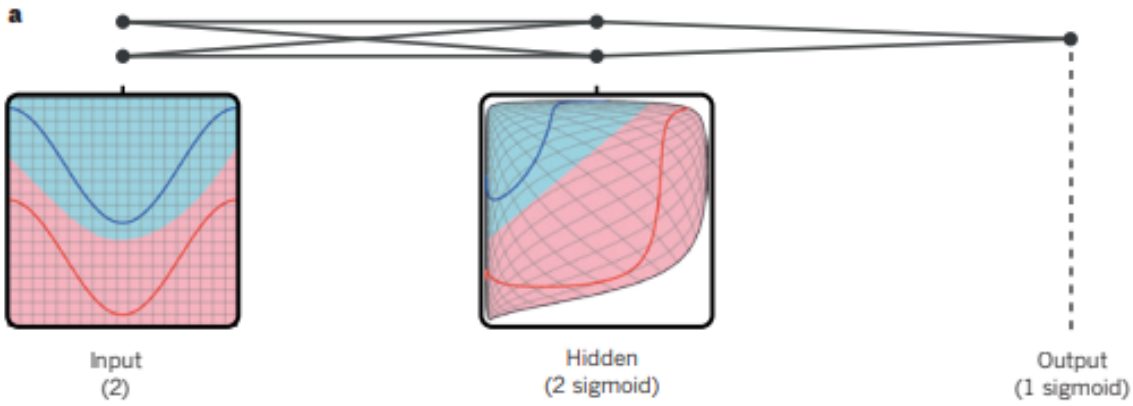

Figure 3.5 Sample deep Neural network

Recurrent Neural Networks:

For our work we are going to be focusing on recurrent neural networks. Recurrent neural networks are a type of artificial neural networks in which we can take the output of one layer and input it again into our model to get a better prediction out of it [43]. The layers of a recurrent neural network can either be all interconnected or partially interconnected, depending on the type of model that we are using as can be seen in figure 6.

The reason we want to use recurrent neural networks is to see if keeping context of the conversation into account influences the results that we get. We can't do this using feed forward neural networks.

The extent to which simple RNNs can keep context into account is quite limited, so let's investigate some types of RNNs that can better.

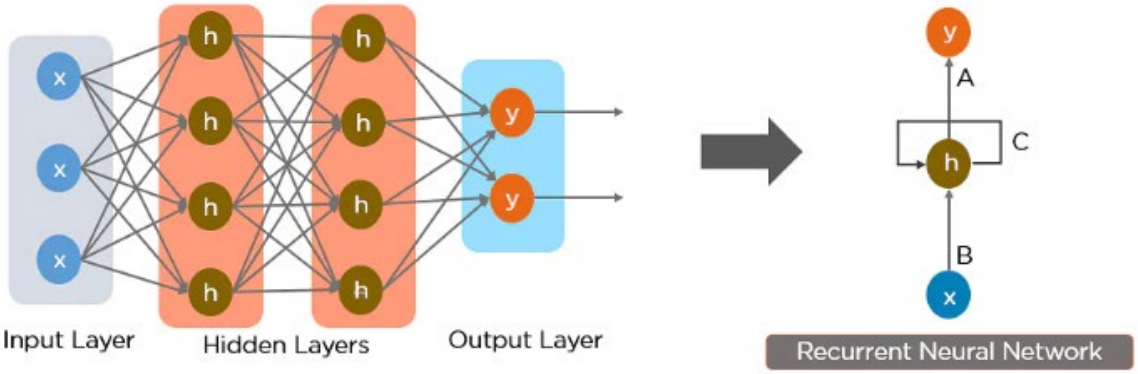

Figure 3.6 Depiction of a recurrent neural network

Types of recurrent neural networks:

There are different types of recurrent neural networks, we are going to be focusing on two that we have used in our research. Long short-term memory [44] and Gated Recurrent Units [45].

Long Short-Term Memory (LSTM) neural network:

LSTM are a type of recurrent neural network that use to process sequential data. What this means is they are used to process data in which information that came before the current datapoint might be important in training the network for the current data.

The distinction that a simple RNN has from an LSTM is that LSTM is a lot more complex, to be more precise it has 3 gates: input, output and forget. in which information can regulate in a better manner. This state is then updated with the new output (that will also be used the output).

The functions of these gates are as follows:

- The **forget gate** is triggered at the very first, it checks the whether the bit is 0 or 1 and decides whether or not it wants to retain information from the previous outputs or not. This is usually done via sigmoid activation.
- The **input gate** is triggered next, this decides whether or not the informat**ion** that is being fed should be stored in the state, this does so by taking a look at the new data as well as the previous data present in the state.

- The **output gate** is triggered at last, which modifies the hidden state by taking into account the new state, the last hidden state as well as the input.

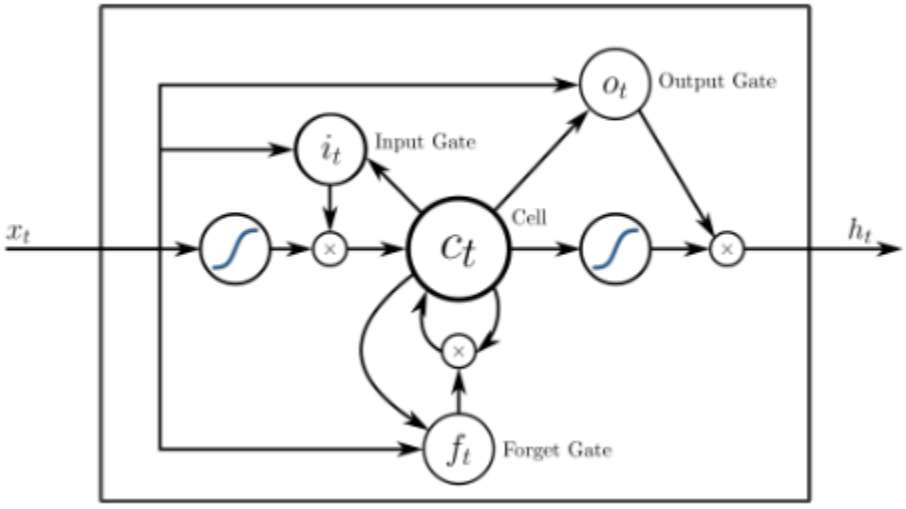

Figure 3.7 Gates of an LSTM

Gated Recurrent Units (GRU):

GRU is a more sophisticated version of Recurrent Neural networks but simpler than LSTM. Simple in a way that while LSTM has 3 gates, GRU has only 2. GRU only has reset and update gate.

- **Update gate** has the same functionality that the output gate has in LSTM, it checks for which and how much information should be retained in the state
- **Reset gate** is synonymous with the forget gate in the LSTM it is used to determine how much information should be forgotten by the neural network.

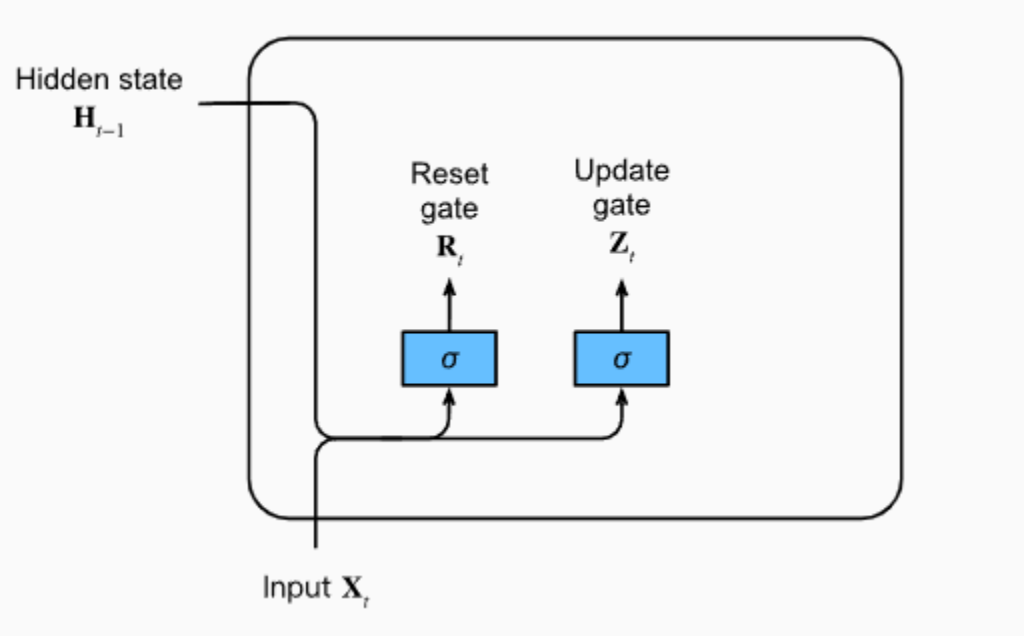

Figure 3.8 Gates in a GRU

GRU has a simpler architecture as compared to LSTM, so understandably it is better suited to smaller datasets. LSTM on the other hand is preferred when are dealing with larger datasets.

3.3 Evaluation Metrics

3.3.1 Area under the curve

To understand area under the curve lets first talk about what Receiver Operating Characteristic curve or ROC [46] is. ROC curve, simply put is the is the curve you get when you plot the true positive rate of your classifier against the false positive rate where:

True positive rate (also known as TPR) is the ratio of correctly predicted true positives our of all the positive predictions and is given as:

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

Where TP is the total number of true positives and FN is the total number of false Negatives

And **false positive rate** (also known as FPR) is the ratio of incorrectly detected positives out of all the negative predictions and is given as:

$$False\ Positive\ Rate = \frac{FP}{FP + TN}$$

Where FP is the total number of False positives and TN are the total number of true negatives.

Area under the curve (AUC) is the area under the ROC. It tells us the degree of separability between the classes and for our classification is calculated by the trapezoidal rule.

The value of AUC ranges between 0 and 1 and should be as far away from 0.5 as possible. Ideally the value should be close to 1, but if we get values really close to 0, that does not necessarily mean that our model is not performing well, we can just invert the labels and get the desired output. An area under the curve of 0.5 means that our model has learnt nothing and cannot differentiate between any of the classes.

3.4 Flow of our experiments

The basic flow of our experiments is as follows:

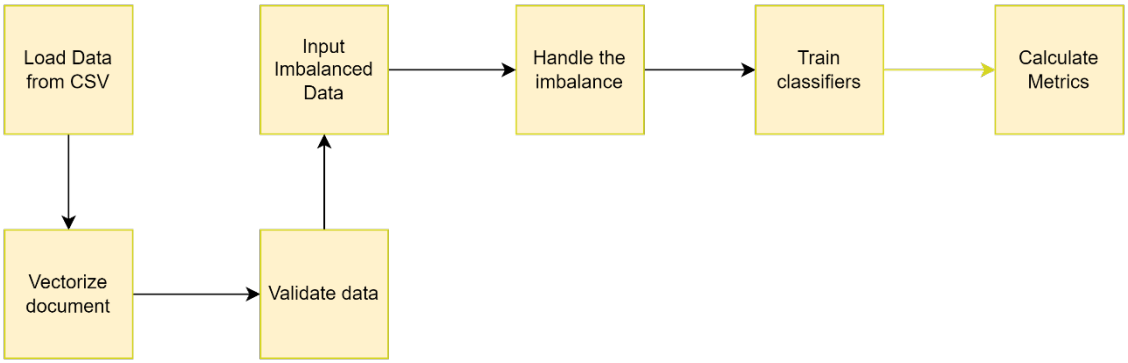

Figure 3.9 Flow of our experiments

We can categorize our experiments in 4 different categories.

- Running Machine learning algorithms on imbalanced data

- Running Machine learning algorithms on balanced data
- Running Deep Neural Networks on imbalanced data
- Running Deep Neural Networks on imbalanced data

Chapter 4

Experiments, results, and discussions

In this chapter we will discuss the experiments we performed, our results and in-depth discussion about what we can deduce from them.

4.1 Experimental setup

All the code was written in python 3.9. The major libraries used, and their reasons are usages are given in table in table:

Library	Version	Usage
Xml	Native	Used to load data from the csv
Pandas	1.4.4	To visualize and organize data
NumPy	1.23.3	To create arrays to feed the data to our models
Genism	4.2.0	To train word2vec as well as all work related to NLP preprocessing
Imblearn	0.10	For applying SMOTE to handle data imbalance
Scikit-learn	1.1.2	All the machine learning models were used from the scikit-learn library
TensorFlow	2.10.0	All the deep learning models were custom created using TensorFlow
Matplotlib	3.6.0	All the graphs created were done so using matplotlib

Table 4.1 Python Libraries used

4.2 Results

The tables in the following sections show the experiments we ran on the balanced and unbalanced datasets using our custom trained word2vec model and why we needed to train it.

4.2.1 Why we trained our word2vec

But before getting into the actual experiments let's spend some time talking about how we had to train our own word2vec and how that affected our models and our results.

We first created vectors using google news word2vec model [36]. It was around 3.5 GB in size, so it had been trained on a lot of data from news outlets, but when we ran it on our models, we got garbage results.

For our Machine Learning models,

Model	AUC (Without balancing)	AUC (With Balancing)
Gaussian Naïve Bayes	0.5	0.5
QDA	0.5	0.5
LDA	0.5	0.5
Logistic Regression	0.5	0.5
Random Forest	0.5	0.5
K-Nearest Neighbor	0.5	0.5

Table 4.2 Machine Learning models trained with vectors from google W2V

For our deep learning models:

Deep learning model	Layer	Units	AUC (Without balancing)	AUC (With Balancing)
Recurrent Neural Network	LSTM	2	0.5	0.5
Recurrent Neural Network	LSTM	32	0.5	0.5
Recurrent Neural Network	GRU	2	0.5	0.5
Recurrent Neural Network	LSTM	32	0.5	0.5

Table 4.3 Deep learning models trained with vectors from google W2V

This might seem peculiar at first but after looking at some of the conversations with a naked eye the problem was quite self-evident.

The Word2vec model that we downloaded from google, is trained Google news data, so it does not account for the huge amount of slang language that is incorporated in millions of online chat conversations.

As can be seen in Figure 6, these conversations heavily involve slang language, language that has use of words like "*luv, u, bby, gurl etc.*". For that reason, any pretrained word2vec model would not suffice in this scenario.

Figure 4.1 Slang language example

4.2.2 Using Machine Learning methods

The performance of our methods using machine learning methods before and after balancing of data is given in table 4.4.

Metrics	Area under the curve (Without balancing)	Area under the curve (With Balancing)
Gaussian Naïve Bayes	0.56	0.53
QDA	0.55	0.79

LDA	0.49	0.41
Logistic Regression	0.50	0.42
Random Forest	0.50	0.93
K-Nearest Neighbor	0.50	0.79

Table 4.4 Performance of Machine Learning methods

4.2.3 Using Deep learning methods

The performance Recurrent Neural network with LSTM layer is given in table 4.5

Layer	Units	Dataset	Input type	Area under the curve
LSTM	2	Unbalanced	Word2vec	0.61
LSTM	2	Balanced	Word2vec	0.64
LSTM	32	Unbalanced	Word2vec	0.69
LSTM	32	Balanced	Word2vec	0.82

Table 4.5 Performance of LSTM

The performance of Recurrent Neural networks with GRU layer is given in table 4.6

Layer	Units	Dataset	Input type	Area under the curve
GRU	2	Unbalanced	Word2vec	0.70

GRU	2	Balanced	Word2vec	0.72
GRU	32	Unbalanced	Word2vec	0.74
GRU	32	Balanced	Word2vec	0.76

Table 4.6 Performance of GRU

4.3 Discussion

We have already deduced in section 4.3.1 that we needed to train our own word2vec model in order to get good results. Table 4.2 and Table 4.3 support that theory as the AUC in all those cases is 0.5 which means that the models we trained were doing random predictions and had learnt nothing.

When we ran the tests on our machine learning models without balancing out the data the results were more than underwhelming as can be seen in table 4.4. Nearly all of our machine learning models have an AUC very close to 0.5 which means that the performance of the models was based on random guesses rather than an intelligent prediction and that the models could not differentiate between a predatory and a non-predatory message.

Deep learning methods on the other hand were considerably more intelligent in their learning approaches as even without balancing out the data we got an AUC of greater than 0.7 for both LSTM and GRU for imbalanced dataset. GRU with 32 units had an AUC of 0.74 performed the best out of all the methods when the data was imbalanced.

It is evident that applying SMOTE had a noticeable difference in the results for pre7dator classification using Machine Learning. With Quadratic discriminant analysis and K-NN giving an AUC of 0.79 and Random Forest outperforming every other Machine Learning and Deep learning method with an AUC of 0.93.

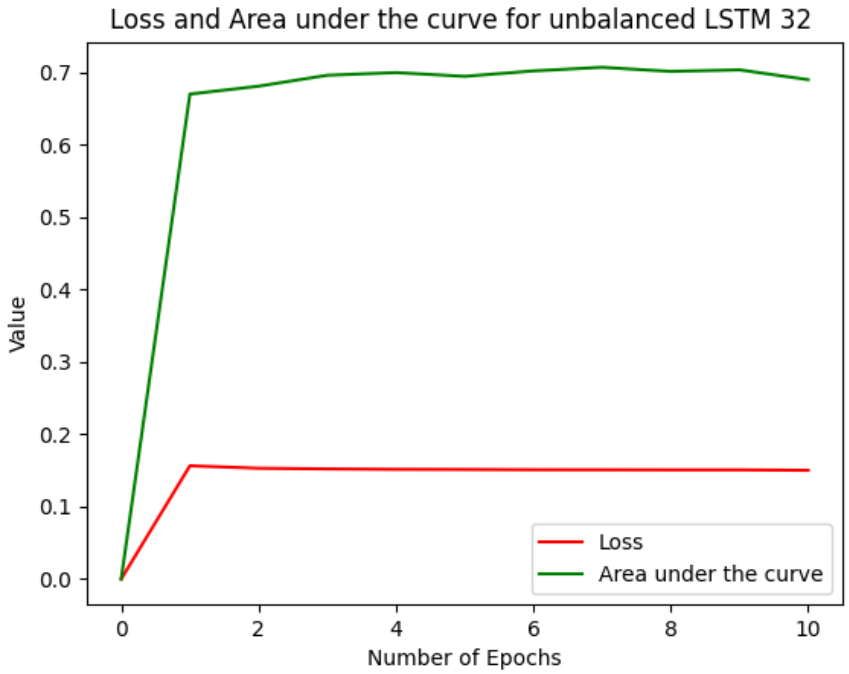

Figure 4.2 Loss and AUC for LSTM with unbalanced data

Figure 4.3 Loss and AUC for LSTM for unbalanced data

Similar to how Machine Learning models gave better results after handling the data imbalance, Deep learning models also improved their performance after applying SMOTE. The increase in performance for GRU was rather underwhelming though. For GRU with 32 units there was an increase in AUC from 0.74 to 0.76, which although an increase, is very minor. LSTM on the other hand saw a very significant performance boost when we applied SMOTE. LSTM with 32 units improved its AUC from 0.69 to 0.82, which was the best result for any of our machine learning models. Figures 4.3 and 4.4 give a very good depiction of how the loss is stagnant after the first couple of epochs in deep learning models when the data is imbalanced, where as the model is consistently learning and improving when working with balanced data.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

We could improve the method of predator detection by handling the data imbalance and proposing area under the curve as a better measure of how the model is performing. We have used synthetic minority oversampling technique to handle the data imbalance and AUC to evaluate our machine learning and deep learning methods. Our experiments showed that handling data imbalance using SMOTE and using that data significantly increased the performance of our models and AUC was a very accurate measure of how good or bad our models were performing.

5.2 Future work

This thesis has laid the foundation for using SMOTE to handle imbalanced predatory datasets and test their performance using AUC, there is a lot of work that can be done upon this. Such as:

- Fine tuning the hyper-parameters of the recurrent neural networks to try and achieve better results. We have fine tuned our models a lot using grid search, but I believe with more time and a faster system, these RNNs can yield even better results.
- Using Bidirectional recurrent neural networks which can take into account, the current, previous as well as the future messages, this can be a good step to increasing the results that we already have.
- Use Google's twitter/social media word2vec and see the results.

These are two of the many ways that our thesis can be expanded upon as there is not much literature on predator detection yet.

Bibliography

[1] S. Araji and D. Finkelhor, "Explanations of pedophilia: Review of empirical research," *Bulletin of the American Academy of Psychiatry and the Law*, 1985.

[2] J. B. Murray, "Psychological Profile of Pedophiles and Child Molesters," *J Psychol*, vol. 134, no. 2, pp. 211–224, Mar. 2000, doi: 10.1080/00223980009600863.

[3] J. A. Fernandez, J. A. Fernandez, and R. Aga Mohd Jaladin, "Beware of the menacing monsters around us: protecting Malaysian children from sexual abuse," *Br J Guid Counc*, pp. 1–10, Jun. 2021, doi: 10.1080/03069885.2021.1938970.

[4] R. O'Connell, "A typology of child cybersexploitation and online grooming practices." 2003.

[5] S. Craven, S. Brown, and E. Gilchrist, "Sexual grooming of children: Review of literature and theoretical considerations," *Journal of sexual aggression*, vol. 12, no. 3, pp. 287–299, 2006.

[6] A. A. Gillespie, "Child protection on the internet-challenges for criminal law," *Child & Fam. LQ*, vol. 14, p. 411, 2002.

[7] G. M. Winters and E. L. Jeglic, "Stages of sexual grooming: Recognizing potentially predatory behaviors of child molesters," *Deviant Behav*, vol. 38, no. 6, pp. 724–733, 2017.

[8] M. L. Williams and K. Hudson, "Public perceptions of internet, familial and localised sexual grooming: Predicting perceived prevalence and safety," *Journal of Sexual Aggression*, vol. 19, no. 2, pp. 218–235, Jul. 2013, doi: 10.1080/13552600.2012.705341.

[9] M. Welner, “Child Sexual Abuse: 6 Stages of Grooming,” *Retrieved from website on August*, vol. 10, 2020.

[10] V. Kashuba, N. Nosova, and Y. Kozlov, “Theoretical and methodological foundations of the physical rehabilitation technology of children 5-6 years old, with functional disorders of the support-motional apparatus,” *Journal of Education, Health and Sport*, vol. 7, no. 4, pp. 975–987, 2017.

[11] J. Dalins, C. Wilson, and M. Carman, “Criminal motivation on the dark web: A categorisation model for law enforcement,” *Digit Investig*, vol. 24, pp. 62–71, Mar. 2018, doi: 10.1016/j.diin.2017.12.003.

[12] H. Amanda, “Policing international child pornography on the dark web,” *Syracuse J. Int’l L. & Com*, p. 353, 2015.

[13] T. Susi, N. Torstensson, and U. Wilhelmsson, “‘ Can you send me a photo?’: A Game-Based Approach for Increasing Young Children’s Risk Awareness to Prevent Online Sexual Grooming,” in *DiGRA 2019, The 12th Digital Games Research Association Conference, Kyoto, Japan, August, 6-10, 2019*, 2019.

[14] G. M. Winters, L. E. Kaylor, and E. L. Jeglic, “Sexual offenders contacting children online: an examination of transcripts of sexual grooming,” *Journal of sexual aggression*, vol. 23, no. 1, pp. 62–76, 2017.

[15] O. G. Omiunu, “Online Sexual Grooming among Female Secondary School Students: A Nigerian Case Study”.

[16] M. Ebrahimi, C. Y. Suen, and O. Ormandjieva, “Detecting predatory conversations in social media by deep Convolutional Neural Networks,” *Digit Investig*, vol. 18, pp. 33–49, Sep. 2016, doi: 10.1016/j.diin.2016.07.001.

[17] C. Morris, “Identifying online sexual predators by svm classification with lexical and behavioral features,” *Master of Science Thesis, University Of Toronto, Canada*, 2013.

[18] A. E. Cano, M. Fernandez, and H. Alani, “Detecting Child Grooming Behaviour Patterns on Social Media,” 2014, pp. 412–427. doi: 10.1007/978-3-319-13734-6_30.

[19] A. Kontostathis, L. Edwards, J. Bayzick, A. Leatherman, and K. Moore, “Comparison of rule-based to human analysis of chat logs,” *communication theory*, vol. 8, no. 2, p. 12, 2009.

[20] N. Pendar, “Toward spotting the pedophile telling victim from predator in text chats,” in *International Conference on Semantic Computing (ICSC 2007)*, 2007, pp. 235–241.

[21] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski, “Learning to identify internet sexual predation,” *International Journal of Electronic Commerce*, vol. 15, no. 3, pp. 103–122, 2011.

[22] C. Peersman, F. Vaassen, V. van Asch, and W. Daelemans, “Conversation Level Constraints on Pedophile Detection in Chat Rooms.,” in *CLEF (Online working notes/labs/workshop)*, 2012, pp. 1–13.

[23] A. Gupta, P. Kumaraguru, and A. Sureka, “Characterizing Pedophile Conversations on the Internet using Online Grooming,” Aug. 2012.

[24] S. Craven, S. Brown, and E. Gilchrist, “Sexual grooming of children: Review of literature and theoretical considerations,” *Journal of Sexual Aggression*, vol. 12, no. 3, pp. 287–299, Nov. 2006, doi: 10.1080/13552600601069414.

[25] P. R. Borj and P. Bours, “Predatory conversation detection,” in *2019 International Conference on Cyber Security for Emerging Technologies (CSET)*, 2019, pp. 1–6.

[26] P. Bours and H. Kulsrud, “Detection of Cyber Grooming in Online Conversation,” in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2019, pp. 1–6. doi: 10.1109/WIFS47025.2019.9035090.

[27] F. E. Gunawan, L. Ashianti, and N. Sekishita, “A simple classifier for detecting online child grooming conversation,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 3, pp. 1239–1248, 2018.

[28] E. Villatoro-Tello, A. Juárez-González, H. J. Escalante, M. Montes-y-Gómez, and L. V. Pineda, “A Two-step Approach for Effective Detection of Misbehaving Users in Chats.,” in *CLEF (Online Working Notes/Labs/Workshop)*, 2012, vol. 1178.

[29] M. A. Wani, N. Agarwal, and P. Bours, “Sexual-predator detection system based on social behavior biometric (SSB) features,” *Procedia Comput Sci*, vol. 189, pp. 116–127, 2021.

[30] E. Alawneh, M. Al-Fawa’reh, M. T. Jafar, and M. al Fayoumi, “Sentiment analysis-based sexual harassment detection using machine learning techniques,” in *2021 international symposium on electronics and smart devices (ISESD)*, 2021, pp. 1–6.

[31] G. Inches and F. Crestani, “Overview of the International Sexual Predator Identification Competition at PAN-2012,” Dec. 2012.

[32] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[33] F. Almeida and G. Xexéo, “Word Embeddings: A Survey,” Jan. 2019.

[34] P. Rodriguez, M. A. Bautista, J. Gonzalez, and S. Escalera, “Beyond one-hot encoding: Lower dimensional target embedding,” *Image Vis Comput*, vol. 75, pp. 21–31, 2018.

[35] Q. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *International conference on machine learning*, 2014, pp. 1188–1196.

[36] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” Jan. 2013.

[37] L. Shi, A. Somnali, and P. Guan, “USING CONTINUOUS BAG OF WORDS,” 2022.

[38] Vikramkumar, V. B, and Trilochan, “Bayes and Naive Bayes Classifier,” Apr. 2014.

[39] A. J. Izenman, “Linear discriminant analysis,” in *Modern multivariate statistical techniques*, Springer, 2013, pp. 237–280.

[40] D. R. Cox, “The regression analysis of binary sequences,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, 1958.

[41] T. K. Ho, “Random decision forests,” in *Proceedings of 3rd international conference on document analysis and recognition*, 1995, vol. 1, pp. 278–282.

[42] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[43] L. R. Medsker and L. C. Jain, “Recurrent neural networks,” *Design and Applications*, vol. 5, pp. 64–67, 2001.

[44] A. Graves, “Long short-term memory,” *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.

[45] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” Dec. 2014.

[46] S. Narkhede, “Understanding auc-roc curve,” *Towards Data Science*, vol. 26, no. 1, pp. 220–227, 2018.

Vita Auctoris

NAME:	Muhammad Moeed Khalid
PLACE OF BIRTH:	Hamilton, ON
YEAR OF BIRTH:	1996
EDUCATION:	Punjab College, Lahore, PK, 2014
	COMSATS University, Lahore, PK, 2019
	University of Windsor, M.Sc., Windsor, ON, 2022