National College of Ireland

Human Rights Violation Detection on Social Media

MSc Research Project
Data Analytics

Yash Rajendra Pilankar
Student ID: x19216858

School of Computing
National College of Ireland

Supervisor: Dr. Rejwanul Haque

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Yash Rajendra Pilankar
Student ID:	x19216858
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Rejwanul Haque
Submission Due Date:	31/01/2022
Project Title:	Human Rights Violation Detection on Social Media
Word Count:	6914
Page Count:	20

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.

Signature:	Yash Rajendra Pilankar
Date:	30th January 2022

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):	

Human Rights Violation Detection on Social Media

Yash Rajendra Pilankar
x19216858

Abstract

Apart from entertainment, Social Media is also used for being vocal for Human Rights, as Activists and other User's shares Post in form of text or Media but are not able to reach audience and Human Rights organizations. As there is a staggering growth in technology, usage of advance techniques like Machine Learning and Deep Learning Models can help to get sentiments of these users for better insights and classification. Hence, the objective of the given study was to detect the post/tweets which are about Human Rights Violation over Social Media which can help peace-keeping organization to monitor real-time situation where human rights are being violated. The study was achieved by implementing different experiments of Machine Learning, Natural Language Processing (NLP) and Deep Learning Models on a Dataset fetched from Twitter using Twitter API and were adjudged based on metrics namely accuracy, sensitivity and specificity.For Machine Learning Models accuracy of 98% was achieved whereas on Deep Learning Model decent accuracy of 75% was obtained.Based on the results, given study was able to classify tweets about Human Rights violation and can be used extended further for future use by different organizations.

1 Introduction

In 21st century, where there is significant growth in technologies, modern ideologies, development in various science and research and growth of every Nation, there are still many locations around the world where people are not availed with basic human rights. In current affairs there are enormous incidents that are befalling in the world where the given human rights are being violated and are gone unnoticed. Many activist around the world try to raise these issues once they are aware about the incident. Likewise, the activist and media reporters try to be on field risking their life for coverage of the news and fighting for the human rights.One of the article in Hindustan Times by Laskar and Sunny (2021) had reported on the recent heartbreaking occurrence of one of India's most renowned and Pulitzer Prize-winning photojournalists, Danish Siddiqui passed away while reporting on an incident where human rights were violated. Since advance technologies are being adopted by different industries similarly, it should be adopted by the peace making organizations for keeping harmony and peace in the world.

There are multiple crowd-sourced technological solutions build around like Ushahidi is one of the technology used for monitoring situation around and tag it live on map. Syria Tracker [1] is one such tool that reports about Human Rights being violated and tag the location that makes neighborhood aware about the event. Similarly, by using

[1] `https://syriatracker.crowdmap.com/`

advance technologies like Natural Language Processing (NLP), Alhelbawy et al. (2020) developed similar platform that monitors abuse of Human Rights. Similarly, the given study focuses over detecting Human Rights Violation.

Over the past few years, there is significant growth in usage of Social Media and by-far people are being vocal about their Human Rights and share their thoughts over it. Since past decade, researchers are investigating more over sentiment analysis by analyzing data from Social Media, but their major objective is to find the mood, behavior and current sentiment of the user. Akin research was conducted by Neethu and Rajasree (2013), to get the sentiments of user but it only focus on to get the mood of the user whereas Waseem and Hovy (2016) were trying to get intentions of the user and classified if the tweets based on hate-speech or not. Likewise, the given study had utilized similar methodology but it focused on "facts" rather than an emotion or intention. For dataset the data was gathered from Social Media (Twitter) as people tend to share about the event and incidents occurring nearby them in form of post/tweet. Since, millions of post on Social Media goes viral over the internet about entertainment that makes voices about human rights being violated getting cornered. If these post and Media reaches their audience it will make nation unite together and will be able to give a human their basic rights which they deserve.

Research Question: *To what extent Machine Learning and Deep Learning can help to detect factual posts about Human Rights Violation over Social Media with the assistance of Natural Language Processing?*

The research presented solution over aforementioned question and achieved the objective which was to detect factual post on social media that addresses over human right abuse. This can help peace-keeping organizations to track and resolve the situation accordingly. The contribution of the given research is not enclosed to human rights but also over the diverse application of Machine Learning and Deep Learning Models and along the understanding of Natural Language Processing that helped models to differentiate factual tweets compared to other tweets. It also addresses over the collection of dataset since, the data gathered was not readily available rather had to be taken using Twitter API [2].

The given report is systematized as follows, section 2 discusses over previous related work conducted by researchers that enables to understand the domain, different methodologies and applications of Machine Learning and Deep Learning Models. The section also shows how Sentiment Analysis is done to understand the intention and opinion of users. Further, section 3 gives details about proposed methodology and section 4 gives overview of the architecture along with the methods utilized whereas section 5 shows about the implementation of different models based on Machine Learning and Deep Learning. Section 6 glimpse over the findings and results are presented and discussed. Lastly, section 7 concludes over the given study and shares the future aspects of the research.

2 Related Work

Over the decade, billions of users have been engaged to Social Media where trillions of Gigabytes data exist on internet. Collecting relevant dataset from these social media is one of the tedious task. Although, many researchers have used web scrapping tools and API's provided by Social Media for gathering data for their dataset. Twitter being

[2] `https://developer.twitter.com/en/docs/twitter-api`

most popular social media and a micro blogging website where users majorly tweet about opinion, news and event multiple researchers have extracted tweets for creating their dataset. Even after collecting the dataset, labeling and annotating the same dataset is difficult. But researchers have used diverse ways for labeling and annotation these tweets where they tried to classify them into Binary Classification or Multi-Class Classification wherein the Binary Classification of dataset was based on "Yes/True" or "No/False" and for Multi-Class Classification it was over "Positive", "Negative" or "Neutral". Further, for getting sentiments and classification researchers had implemented Machine Learning and Deep Learning and used various metrics for adjudging their model. The given section covers diverse approach and methodologies along with different technologies used by these researcher and check if their objective has met where in section 2.1 gives overview of how intention of user are analyzed using sentiment analysis whereas section 2.2 indicates about how researchers were able classify sentiments of user's tweet over a certain event and section 2.3 gives comparison over the existing methodology that helps to Detect/Monitor Human Rights Violation.

2.1 Identification of User Intention using Sentiment Analysis

Earlier in the given decade where machine learning techniques were getting popular, there was curiosity in companies to know moreover how customers react to their product. However, Neethu and Rajasree (2013) took this as challenge and conducted research over Sentiment Analysis on customers by analyzing Tweets about electronic products using Machine Learning approach. The motive was to get the opinion of the users over the product which were classified in "Positive" and "Negative". For creating dataset they developed personalized Feature Vectors and were able to get meaningful tweets which were annotated and given to various classifier like Naïve Bayes and Support Vector Machine (SVM) had mind-boggling accuracy of around 90%. The approach taken by above paper shows that Machine Learning technique are effective over classifying the sentiments of users. Further for the given domain multiple research have been conducted and researchers are keen to deep dive and get intentions of a user / customer /person.One of such research was conducted by Purohit et al. (2015) for getting the intentions like if person is "seeking help", "willing to give help" or else "none". They had challenges like ambiguity in interpreting data and sparsity over relevant behavior of user in social dataset. They created a hybrid model using knowledge-guided patterns in top-down approach and bag-of-tokens in bottom-up approach. They also included pattern-set for gaining knowledge from various sources like psycho-linguistics to tackle the ambiguity challenge, get the social behavior of the user along with contrast patterns to tackle the challenge sparsity. Post applying the techniques they has promising results which was gain of 7% from the base model. Another approach of feature extraction and label weighting was taken by Hamdan et al. (2015) for performing sentiment analysis. In their methodology they included multiple characteristics for feature extraction like usage of polarity score for different lexicons that had 4 manual and 6 automated lexicons. They also included customized slang dictionary of Twitter. As dataset being imbalanced the researchers used weighting schema that adapts negative and positive labels that helped their models to understand the bias of the class. For classification, Logistic Regression algorithm was applied that gave F1-score of 64% which was acceptable but the model would have been improved if feature extraction methods like emoticons and hashtag expressions as positive or negative would have been included as it also tends to give weightage for the

post.Similarly, Zahoor and Rohilla (2020) performed research of Sentiment Analysis using Lexical and Rule based approached over Twitter dataset. Further, for sentiment analysis they used NLP libraries i.e. TextBlob and VADER that gave sentiments of the text. The research just focused over getting sentiments whereas missed out on usage of Machine Learning approaches. On the other hand, Zimbra et al. (2018) conducted review of 28 Information Systems that are utilized for conducting sentiment analysis. The researchers collected Tweets that were bifurcated in different industries like Pharma, Retail, Security, Technology and Telecom and further sentiments were classified "Positive", "Negative" or "Neutral". The given dataset was further given as input these various Information Systems where it gave an average accuracy of 70%.

Identification of user intent will always be an advantage for company where they can get the cluster of potential customer and also for Media platforms to throw ads as per customer behaviour and understanding. One such research was accomplished by Haque et al. (2019) where the objective was to find user intent and Deep Learning technique like Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM) was utilized and F1-score of 83% was obtained. Similarly, another study by Singh et al. (2020) focused on identification of complaints in Hindi as data was collected by crawling Amazon website and Youtube where state-of-art methods were applied for classification that fetched accuracy around 68%. Further Twitter dataset was also considered and Deep Learning techniques were applied. The above research shows real application of Machine Learning and Deep Learning where it can classify Human intentions.

A comparative study of machine learning algorithms were conducted by Kanakaraddi et al. (2020)to get the opinionated sentiments of users. The researchers used stream of data fetched using Twitter API and performed feature extraction for labeling the tweet as positive and negative sentiment. As being comparative study they used Naive Bayes, Support Vector Machine, LSTM, CNN, Random forest, Max Entropy among which SVM had better accuracy of around 79%. The study showed that Machine Learning models are capable of classifying the sentiments. Likewise, Mandloi and Patel (2020) also showed a comparative study but they had made improvisation by using Natural Language Processing (NLP) which seemed to be missing in the previous papers. They used Tokenization technique along with other feature extraction techniques like Parts of Speech, usage of Polarity for getting sentiment label of the context and removal of stop words and Non-English words. These methods tweaked in the model development helped to improvise the results where the accuracy was above 82%. The experiment performed by Lim et al. (2020) showed the usage of other NLP techniques like Term Frequency — Inverse Document Frequency (TF-IDF), Word2Vec and Embedding from Language Models (ELMo). The researchers also criticised over the usage of Parts-of-Speech technique as it is not suitable for micro–blog text whereas they used Stemming by using SnowballStemmer also applied Tokenization for getting the segments of the text using Tokenizer function in Keras. The results obtained were near by 80% and suggests that Natural Language Processing techniques are better suitable for Micro-Blog text like Tweets of Twitter as it understands the context and content of the text.

Researchers have not just limited the domain of collecting the sentiments for companies and industries but have also explored the Social cause as many people are getting trolled, bullied and get threats over social media. One of such research was conducted by Febriana and Budiarto (2019) to get the intentions of the user where they collected Twitter dataset for 2019 when elections were being conducted and many cases of cyber-bullying and hate speech were seen. As the standard pre-processing steps were applied

for Tokenization, Latent Dirichlet Allocation (LDA) was used. For sentiments polarity of the context was calculated which help to bifurcate tweets between positive and negative. Similarly, an experiment was performed by Pandey et al. (2018) where they used Distributional Semantics Approach for detecting intent of the user in conversation over Sexual Assault where they categorized into Accusational, Validational, or Sensational intent. The annotation of data was conducted based on the confidence score of the text for a given label. For feature extraction s bag-of-words, ngrams, and Part-of-Speech tags were utilized. Further Word2Vec was used for creating vector of text and was fed to Linear Model and Convolution Neural Network (CNN). The given study suggests for getting intention Natural Language Processing techniques are suitable for getting the context as it understands the context but also Vectorization techniques like TF-IDF, Word2Vec or LDA are likely to be utilized more often.

2.2 Classification of Sentiments Over Event-based Post on Social Media

An Event is an occurrence of an occasion and getting context of such events is difficult over Social Media information gets viral easily. Collection of such event based dataset is tedious as there includes noisy data and such dataset takes time for gathering and creation. These datasets were gather by different means and processed accordingly. A research was accomplished by Patil and Chavan (2018) to get sentiments of user and also detect the event. The researchers proposed SegAnalysis framework that will segment the Tweets using Parts-of-Speech and further classified by Naïve Bayes and used online clustering for detection of events in a Tweet. Another side of Event-Based sentiment analysis is News based sentiments where Word Emotion Association Network (WEAN) technique was proposed by Jiang et al. (2017) as it challenge the existing system and supervised learning techniques. The dataset was obtained from Sina Microblog about the event of "The Malaysia Airlines MH370". Based on the application of WEAN an average accuracy of around 78% was obtained. Similarly, Shahare (2017) performed relatively same analysis by using Naïve Bayes and Levenshtein algorithm which had an accuracy over 78%.

One of the popular event of the world was the 2016 elections in USA where Donald Trump had won them where data played an important role for the turn around. The given study was examine by Somula et al. (2020) where they took dataset from Twitter and were performing prediction of election based on the sentiments of the users they had an accuracy of 62% where they predicted 31 states winner correctly out of 50 states. Similarly, another event of "Khan Shaykhun Syria Chemical Attack" was investigated by Bashir et al. (2021) where 13,156 tweets over the event were collected using Twitter API and were annotated using Orange Data Mining Software for sentiment analysis. But the limitation of such experiment is just to classify sentiments but doesn't focus over application of state-of-art and unsupervised machine learning techniques.

2.3 Detect or Monitor the Media over Human Rights Violation

Violation of Human Rights is a different content where the context of the text has invasive meaning but compared to Hate Speech and Offensive sentiments. One of such study was conducted by students Kalliatakis et al. (2017), to detect human rights violation using image based dataset which were categorized and annotated manually as child labour,

violence performed by police, child soldiers, refugees and normal images. They used ConvNet as pre-trained model and applied different models based on CNN architecture. The constraint in the research was about the size of the dataset and since images were taken form search engine they may or may not be clear over the factual occurrence of such event.

Further, another research was performed by Azizan and Aziz (2019) for detection of terrorism where they traced a user's account and fetched all the tweets and analysed the sentiment of a tweet i.e. positive/negative/neutral and compared the same with previous tweet's sentiment to conclude over if the account leads to terrorism or not. Based on previous research work, they used Naïve Bayes since it has faster classification rate over Support Vector Machine (SVM) that gives better accuracy compared to Naïve Bayes. But the research showed unclear results and were dependent on previous researcher's conclusion. On the contrast, Ahmad et al. (2019) performed similar experiment, over getting Tweets that are related to propagation of terrorism on Twitter based on sentiments were classified as extremist or not extremist where they used Deep Learning Model that was combination of LSTM +CNN where the output from LSTM was fed into CNN. Based on various experiments performed by researchers the best F1-score of around 88% was obtained that gave strong verdict over the usage of Deep learning models which were opposed by Azizan and Aziz (2019).

Another study by Fitri et al. (2019) showed that how Machine Learning are still effective in text classification where they classified user Sentiments using Naïve Bayes, Decision Tree, and Random Forest Algorithms. The objective was to get sentiments of the users based on Anti-LGBT campaign in Indonesia. The intention of using Machine Learning algorithm was achieved but it majorly classified Neutral sentiments compared to Negative and Positive. Also research missed out over how the dataset was annotated. For monitoring abuse of Human Rights in Iraq Alhelbawy et al. (2020) developed a system called Ceasefire in which users can anonymously report the location where human rights are being violated. The Portal also had additional feature where it automatically fetched tweets regarding human rights violation from Twitter and tagged the location accordingly without disclosing user's information. As the website was used in Iraq, the data being collected was in Arabic hence, Twitter Arabic dataset was taken and pre-processed using Tokenization, Diacritised, Lemma and Stemming were used. For Feature extraction Bag-of-Words and TF-IDF were utilized for weighting scheme. Further Machine Learning and Deep learning algorithms were applied and results were obtained accordingly. The research can be improvised if Pre-trained models over Arabic dataset can be considered.

Further, below given table 1 shows comparison of the existing research fall under the domain of Human Rights violation.

2.4 Summary of Related Work

Based on the above it can be seen that section 2.1 gives overall knowledge of domain of sentiment analysis and getting intentions of users. It also gave overview of Natural Language Processing and Vectorization techniques like Term Frequency — Inverse Document Frequency (TF-IDF) Lim et al. (2020) Section 2.2 emphasis on different applications of Machine Learning and Deep learning algorithm for the given domain and shows limitations of the previous implementation. Finally, section 2.3 shows application of supervised learning techniques for detection of Human rights violation.As there were drawbacks in current implementation and limited exploration over the current topic gave rational to

Author	Dataset	Method	Results
Kalliatakis et al. (2017)	Image-based Dataset from Bing	ResNet 50, ResNet 101, ResNet 152, GoogLeNet, VGG 16, VGG 19, VGG – M, VGG – S, VGG – F, 8-layer Places	Mean Average Precision was around 78% (70/30 train-test split) and around 88% (50/50 train-test split) for VGG–S
Azizan and Aziz (2019)	Twitter Dataset	Naïve Bayes	Similar score gained based on previous research Naïve Bayes.
Ahmad et al. (2019)	Twitter Dataset	LSTM + CNN	F1- Score of 88%
Fitri et al. (2019)	Twitter Dataset	Naïve Bayes, Decision Tree, and Random Forest Algorithm	Best accuracy was over 86%
Alhelbawy et al. (2020)	Twitter Dataset	Naïve Bayes, Linear and Gaussian SVM, CNN, LSTM	F1-score of 75%.

Table 1: Comparison of existing Methodology

this research paper which shows improvised and distinctive methodology that has been discussion in further sections.

3 Methodology

The research aims to detect and report tweets that are related to Human Rights Violation so that it will help NGO's, Government or any peace keeping organization to monitor the situation. The steps followed for the given research has been described in the given section and has been applied accordingly. The methodology for the given study was inspired by combination of KDD and CRISP-DM. The given figure 1 shows the gist of the methodology applied in the given research.

- **Business Understanding:** Before any implementation it was necessary to understand the project, domain and should gain knowledge for the same. The first step in methodology was about business understanding which was gained from literature review that shows the overview of sentiment analysis and also gives gist about limitation and drawback in previous papers.
- **Data Gathering:** The data was acquired using Twitter API where it was gathered in two steps, firstly tweets were retrieved from the Twitter user accounts that were

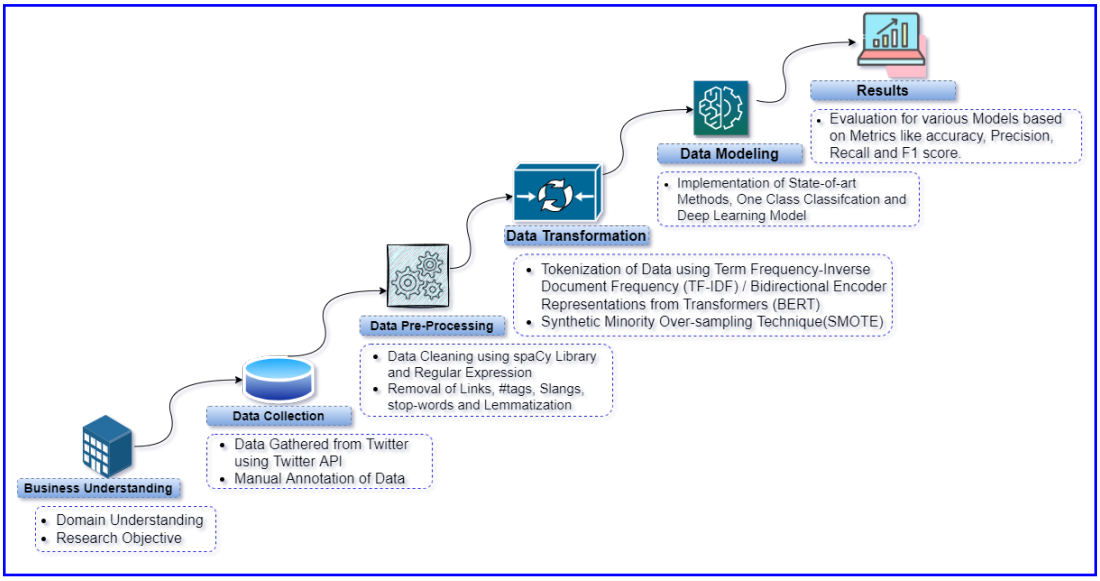


Figure 1: Methodology of the Research

of NGO, Peace-keeping Organization or any Human Rights activist. Secondly, more data was collected using advance search of Twitter API where data filter was applied and further annotated manually by means of human understanding.

- **Data Pre-processing:** Since data is textual, the Tweets were being classified if they are factual Tweets about human rights violation or not. Hence, pre-processing was one of the crucial step as Tweet contains abbreviations, slang's, #tags, links, User tags, etc. Therefore, SpaCy library was utilized which is Natural Language Processing (NLP) based that has functionality for clean the text and remove the irrelevant data in the text.

- **Data Transformation:** The given text-based data was further transformed into tokens using tokenization technique like Term Frequency-Inverse Document Frequency (TF-IDF) where TF-IDF Vectorizer was used to generate numerical statistics and fed to Machine Learning technique whereas Bidirectional Encoder Representations from Transformers (BERT) is transformer based technique which calls over pre-trained model and understands the context of the text and gives encoded vector which was fed to Deep Learning models. Since, data was imbalance Synthetic Minority Over-sampling Technique(SMOTE) was applied on TF-IDF Vector for oversampling the minority class.

- **Data modeling and Results:** Model building was carried where different models were applied over the data gathered and further transformed to be given as an input. Various experiments were carried out where the cleaned and transformed data by TF-IDF and SMOTE was fed to Machine Learning models like Random Forest Classifier, and Linear-based Support Vector Machine for classification of tweets. Also, an experiment was conduct for One-Class Classification where the TF-IDF vector was given as input to One Class Support Vector Machine. Further, data transformed using BERT encoder was fed to Deep Learning Model i.e. Functional Neural Network. The given models were evaluated based on the metrics like accuracy, precision, recall and F1-score. The evaluation of model has been discussed in

the further section.

4 Design Specification

The framework architecture for the given implementation of detection of Human Rights Violation post was designed in 3-Tier architecture. The design specification shows the technologies used for the implementation of the project which characterized in 3 layers i.e. Data Layer, Logical Layer and Client Layer.

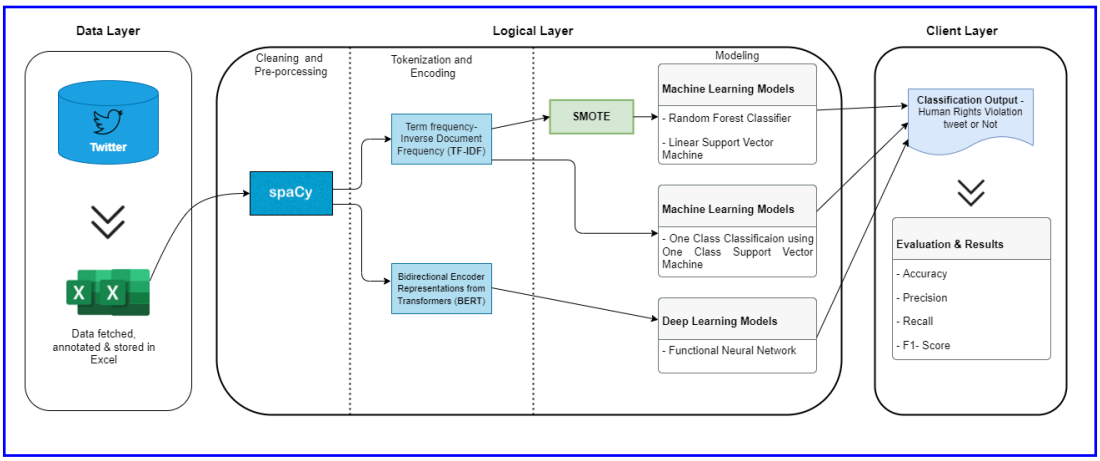


Figure 2: Project Architecture

As seen in the figure 2, the first layer describes over the data layers that includes source of the data from which it was fetched, annotated manually and stored. The logical layer focuses over the transformation of data. It shows the flows from data pre-processing where spaCy library was utilized to gain clean text and was further given for Tokenization and encoding to TF-IDF and BERT. Below is the equation for TF-IDF.

$$W_{x,y} = tf_{x,y} \times log(\frac{N}{df_x})$$

Term Frequency-Inverse Document Frequency (TF-IDF) is a mechanism for getting numerical statistics of a textual data which evaluates how a particular word is relevant in the document calculated by multiplying the term frequency where it focuses on the how many times the word is appearing in a document and inverse of document frequency focuses on the word occurred in the set of documents. The utilization of the given technique was inspired from the work of Alhelbawy et al. (2020) shown in the literature review.

The given research also focuses on the application of Bidirectional Encoder Representations from Transformers (BERT) as the given technique is an upcoming NLP-based model pre-trained model proposed by Google researcher that emphasis on the context of a textual data and gets the encoded values accordingly. An encoder-decoder network that employs self-attention on the encoder side and attention on the decoder side is known as a transformer architecture and BERT is build on such architecture.

The given figure 3 [3], shows the architecture of BERT model. The base model of BERT has 12 layers of encoder stack and also has a large feed-forward network of about 768

[3] `https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/`

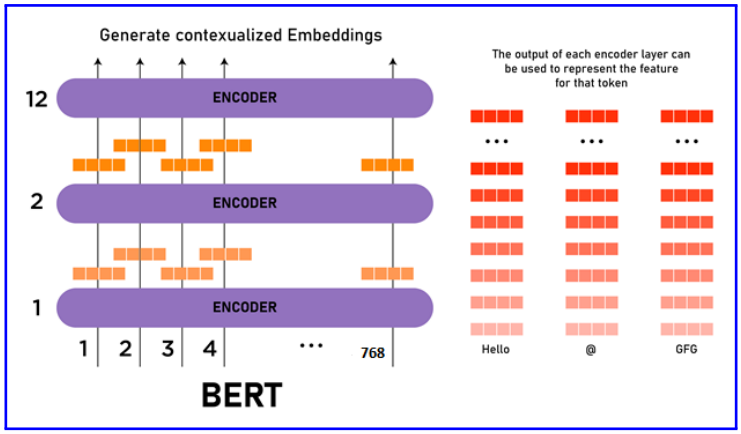


Figure 3: Project Architecture

hidden units. The BERT model gives an output vector of 768 size for BERT Base model. Both encoded text were further given as an input to the Supervise Learning technique by performing train-test split of 75-25 respectively.

5 Implementation

The given project was implemented following the steps from business understand, data gathering towards final output of classification for Human Rights Violation. The project was implemented using Python 3.7 the data was fetched using Twitter API and further accessed locally post data collection, and was imported in Jupyter Notebook where the different experimental models were developed. Before model building the data was cleaned and pre-processed using spaCy library which is NLP based that understands the context of text and gives results accordingly. Since, multiple experiments were performed on the local Machine that has imported Tensorflow and PyTorch along with Sci-Kit Learning libraries that helps in model building. Also, numpy, pandas, matplotlib, seaborn libraries were utilized for data exploration and visualization The local machine is powered by 12GB of RAM, Intel(R) Core(TM) i5-5200U CPU @ 2.20GHz and 2GB of Graphics provided by Nvidia 920M. Hence, the training time for few models was a bit high specifically for Deep Learning Models.

5.1 Data Collection

Data gathering was one of the tedious task in the given research since no ready dataset was available. Getting relevant data was still tedious as it was diving into ocean of data as Twitter generates millions of Tweet daily. Hence, dataset for the given research was gathered using Twitter API accessed using Tweepy [4] library where advance queries were applied. Also, relevant Tweet for the given research would have structure like "Taliban forces unlawfully killed 13 ethnic Hazaras, including a 17-year-old girl, in Afghanistan's Daykundi province after members of the security forces of the former government surrendered" where it states over the violation of Human Rights was noted and the given Tweet was "fact" and not any "opinion" or relevant sentiment. Hence, Fetching data was divided in two parts where first the dataset was collected based on Twitter Accounts of

[4] `https://docs.tweepy.org/en/stable/api.html`

NGOs and Peace keeping Organization like, Amnesty, Human Rights Watch, and Human Rights activist like Malala, Nadia Murad, etc. since, these accounts post relevant factual Tweet about any violation of Human Rights. Still the data gathered was not sufficient, hence the second way for collection of data was by using search function of Twitter API where advance queries were applied. Different filters like language was set to English since as data fetch had tweets in other language another filter of removing Retweets was applied as it would have created redundant data for the model. Queries were fired based on the keywords like, "Child Abuse", "Education ban", "blast", "Genocide", "attack on civilians", etc. were applied to the search function. The data was further manually annotated as factual tweets were required to be classified as Human Rights Violation and the corpus for the given research was created.

5.2 Model Building Approach

Since, the data was in text form it had to be tokenized and encoded there for TF-IDF technique was utilized generating numerical encoded values for the text.

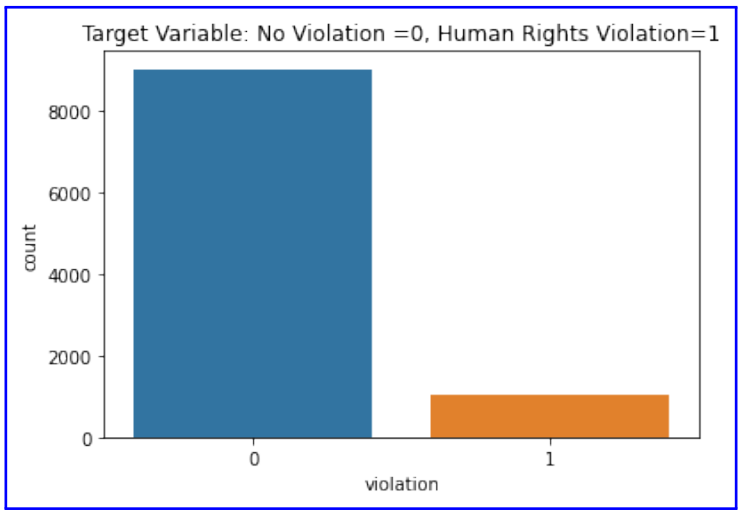


Figure 4: Imbalance class Distribution

As, traditional Machine Learning methods were utilized and dataset was imbalanced as seen in figure 4 it was synthetically up sampled by using SMOTE over minority class and was fed to Random Forest Classifier and Support Vector Machine whereas one of the experiment focused on One Class Classification which is a hot topic since a lot of datasets are largely imbalanced therefore, this technique comes into play where it focuses on majority class and leaves out minority class considering it as outlier. One such experiment was conducted using One Class Support Vector Machine that was trained over majority class. Further the research also explored the Deep learning models where the Data was trained over the layer of pre-trained model of BERT that gave encoded values for a given text based on the context. The model was trained on imbalanced original dataset since BERT was self-sufficient to understand the class weight and context of the text. It was further fed to Deep Learning Models i.e. Functional Neural Network for training the model with the base BERT pre-trained model.

6 Evaluation and Results

In the given research, three different experiments were performed that are described in the further subsections. The experiment 1, in section 6.1 was performed for classification using Machine Learning Techniques (Random Forest Classifier and Support Vector Machine). The experiment 2, in section 6.2 was conducted for classification using One Class Classifier Model (One Class Support Vector Machine). Lastly, experiment 3 in section 6.3 shows classification using Deep Learning Model using BERT as pre-trained model. Post model implementation, they were evaluated and results were analyzed and discussed based on the performance of model. The metrics used for the evaluation were Accuracy which is a performance indicator, along with precision and recall which represents class-wise performance. Also, F1-score were also obtained for overall model performance. All the experiments were conducted on the local machine where the setup was discussed in previous section.

6.1 Experiment 1 – Classification of Human Rights Violation using Machine Learning Models

The aim of the experiment was to train the dataset on machine learning algorithms and evaluate the results based on the classification. Random Forest Classifier and Support Vector Machine models were applied over the cleaned dataset. Since, these technique requires numeric data and the dataset collected was text based. Therefore, TF-IDF was utilized for getting numeric statistic of the text and since dataset was imbalance SMOTE methodology was adopted for up-sampling minority class that made the size of dataset to 18040. The dataset was further split in Train-Test of 75-25 respectively with shuffle set True.

6.1.1 Random Forest Classifier on Human Rights Violation Dataset

The data was suitably transformed and was ready for input to the Random Forest Classifier which is an ensemble based model where it ensembles Decision Tree and naturally extends over Bagging Technique. For Hyper-parameter of Random Forest, GridSearchCV a cross validation technique was applied which is a fit and score method that helps to get the n-estimator parameter for Random Forest Classifier. The criterion was set to entropy as it helps to calculate information gain by the model. Once the parameters were set the Model was build and evaluated based on the metrics and results were obtained were promising as it accuracy and F1-score of the model was 98%. The given Figure 5 shows the results obtained by Random Forest based on hyper-parameter tuning. The model was able to classify Tweets of Human Rights Violated accurately where it missed out over 47 inputs and classified as not about Human Rights Violation.

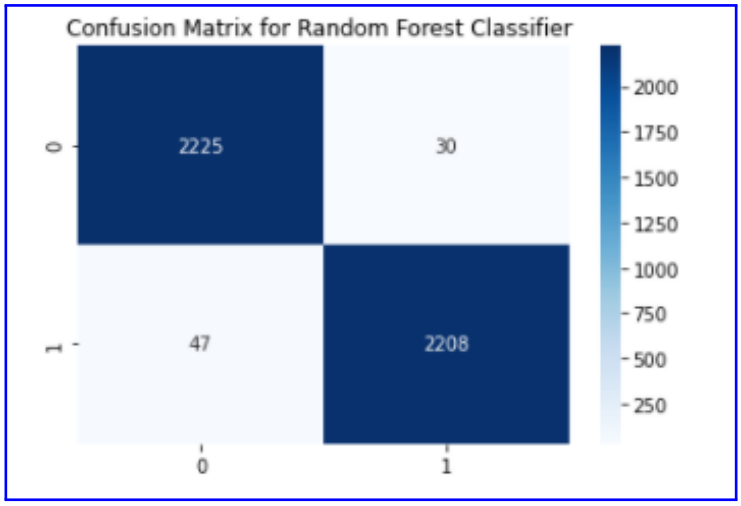


Figure 5: Confusion Matrix of Random Forest

6.1.2 Support Vector Machine on Human Rights Violation Dataset

Similarly, Support Vector Classifier was build which based on Support Vector Machine that provides best fit based on the data provided as input. The model was built upon default values where kernel considered was 'rbf' i.e. Radial Basis Function in which the value is depended on the distance from the origin or from some point.

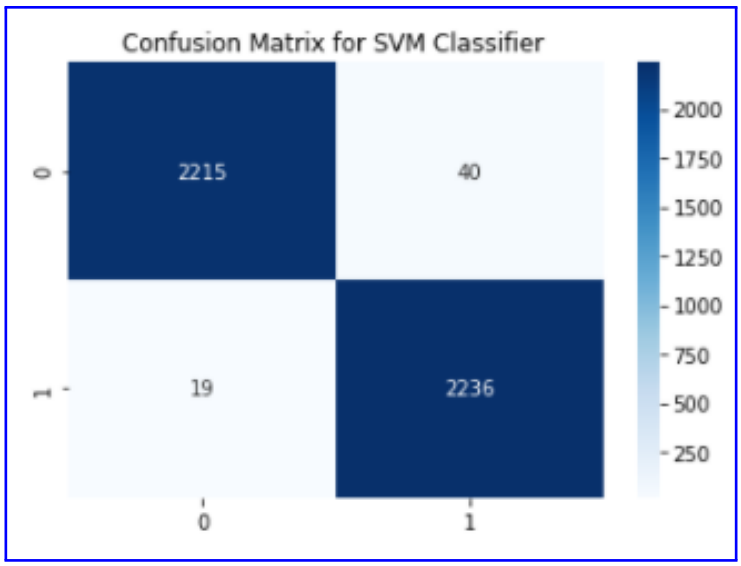


Figure 6: Confusion Metrics of SVM

The given Figure 6 , shows the results based on the metrics and glimpse that the model was precise in classifying the Tweets as it got accuracy and F1-Score of 99%.

6.2 Experiment 2 – One Class Classification over Human Rights Violation

One class classification is upcoming technique utilized for imbalanced dataset where earlier it was used for detecting outlier and anomalies. It is an unsupervised learning method, which attempts to classify only one class and eliminates other class. Apart from

Binary Classification, Support Vector Machine has an extension over the one class classification too. The approach captures the majority class's density and classifies outliers as instances at the extremes of the density function. The given modification in SVM is referred as One Class SVM. As the classifier deals with majority class there for it was set to 1 and the minority class was set as -1.

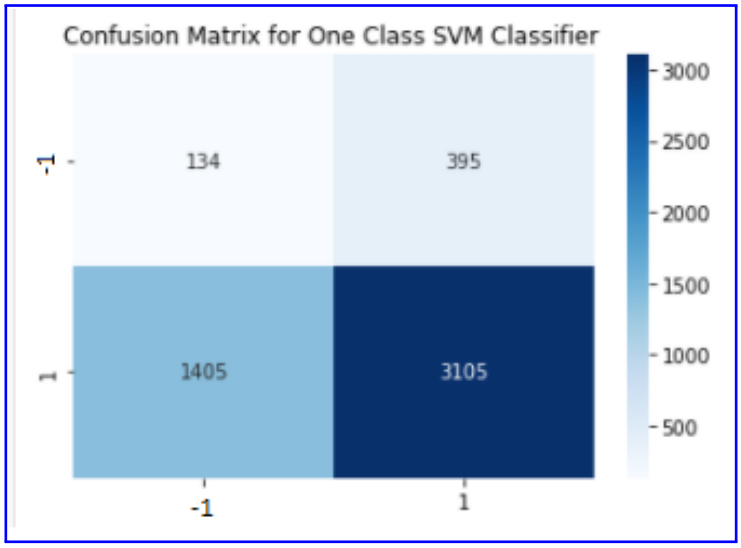


Figure 7: Confusion Metrics of One Class SVM Classifier

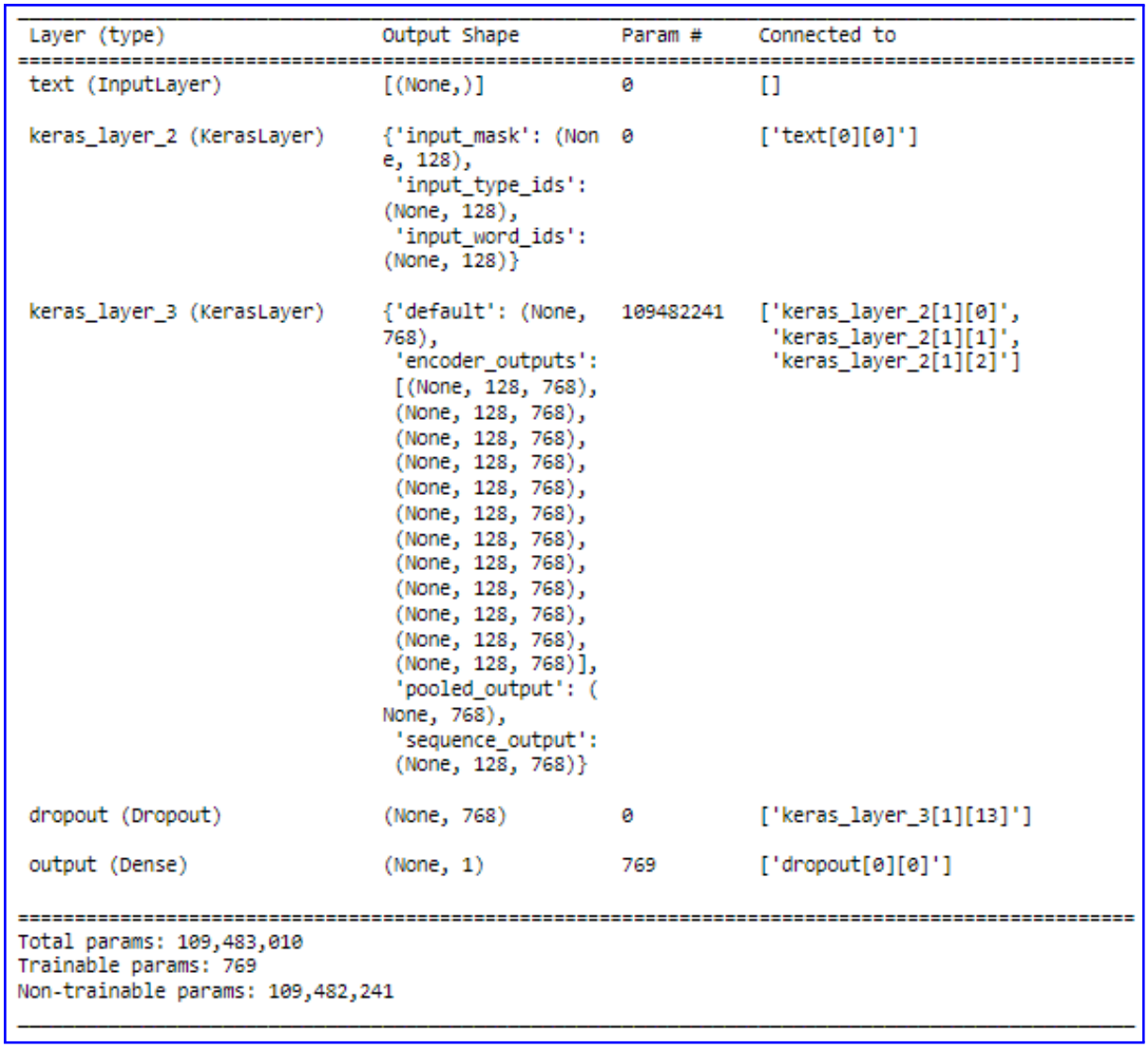
As seen by the results in Figure 7 the accuracy obtained was 64% but taking scores for the class minority which was 13% resulting in worst classification. Based on the metrics it can be evaluated that One Class Classification is not suitable for text based dataset with imbalanced dataset.

6.3 Experiment 3 – Classification using Deep Learning Models.

In the given experiment the dataset was trained over a pre-trained model of BERT which was BERT base model of uncased 12 encoders. The textual data was given as input and an encoded vector of size 768 came as output. The imbalanced dataset was fed to BERT to generate encoded Vector as BERT is sufficient to understand the class weight and it classifies the context of the textual data. Further the Vectorized data was given as input to Neural Network based model ie. Functional Neural Network.

6.3.1 Functional Neural Network on Human Rights Dataset

Functional Neural Network was build on layers as it can be seen in figure 8.



Layer (type)	Output Shape	Param #	Connected to
text (InputLayer)	[(None,)]	0	[]
keras_layer_2 (KerasLayer)	{'input_mask': (None, 128), 'input_type_ids': (None, 128), 'input_word_ids': (None, 128)}	0	['text[0][0]']
keras_layer_3 (KerasLayer)	{'default': (None, 768), 'encoder_outputs': [(None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768), (None, 128, 768)], 'pooled_output': (None, 768), 'sequence_output': (None, 128, 768)}	109482241	['keras_layer_2[1][0]', 'keras_layer_2[1][1]', 'keras_layer_2[1][2]']
dropout (Dropout)	(None, 768)	0	['keras_layer_3[1][13]']
output (Dense)	(None, 1)	769	['dropout[0][0]']

Total params: 109,483,010
Trainable params: 769
Non-trainable params: 109,482,241

Figure 8: BERT Model Summary

It was build on layers where first is input layer, second layer and third layer were hidden layer, fourth Layer was dropout layer and final layer of dense was the output layer. The activation parameter was set to sigmoid as it was a binary classification problem. The Pooled output from the BERT Vector was sent as an input to the neural networks. The parameter tuning was set where Adam was used as optimizer and loss set to binary_crossentropy. Also class weights were calculated and were given to the model to know the imbalance of classes. The model was trained for 10 epoch where the train-test split was of 75-25 respectively. The given Figure 8 show the summary of the model created with the suitable parameters and was fined tuned accordingly. As the model was trained over 10 epochs the loss gradually improved where on first epoch the values were loss: 0.6450 - accuracy: 0.6193 - precision: 0.1621 - recall: 0.630 where on every iteration the model improvised and gained scored of loss: 0.4520 - accuracy: 0.7888 - precision: 0.3082 - recall: 0.813. The model was further evaluated based on Test data. Based on the Results as shown in the Figure 9 it can be seen that the model had Decently performed and was able to get an accuracy of 75%. The model took the consideration of the weights of class and has classified the good amount of samples. From the confusion matrix it can be seen that it missed 29 cases which is considerable but it classified 592 cases of class

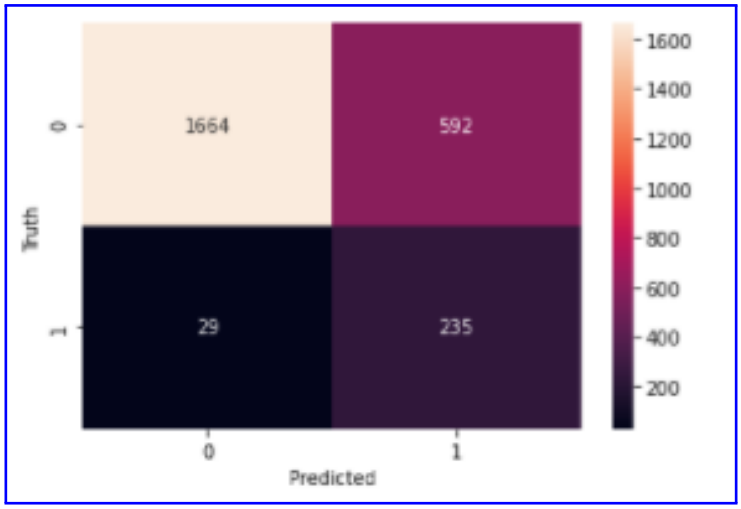


Figure 9: Confusion Metrics of BERT

No as Yes this indicates garbage / Non Violent Tweets were included which harms the precision score of Class 1.

Summary Evaluation and Results: Every experiment were constructed and moulded differently based on the dataset and model. The experiments had same dataset but was resized based on the requirement of the model where in Machine Learning models the dataset was over-sampled using SMOTE as it can improvise the models performance and was able to remove the bias whereas other experiments were made to handle the imbalanced dataset hence, they where tested with the original imbalanced dataset. Further in the Section 6.4 , the overall comparison of the experiments has been discussed.

6.4 Discussion

Based on the knowledge, limitations and recommendation discussed in the literature review, the given research was able to implement the tweaked metholodology that can detect the Tweets about Human Rigths Violation. The this study, three experiments were being conducted which had same same dataset which was moulded according to the ability of the model. All the experiment had same aim but different paths were taken over the journey of development of the models based on the requirements and notions of different model. The experiment also showed utilization of different Tokenization and encoder techniques i.e. TF-IDF and BERT. The dataset was also splited into train-test with ratio of 75-25 and was fed to various models and were evaluated based on the metrics that is discussed below.

The given Table 2 shows the comparison of all the experiments conducted and were evaluated based on the metrics accordlingly. The results are displayed for the Positve class i.e. Tweets are about Human Rights Violation as the objective of the given research was to find the same.

Based on the comparison table it can be seen that Machine Learning Models had outperformed other models and key to their better results was the balancing both the class as SMOTE technique was utilzed for oversampling the minority class where as on imbalanced dataset BERT performed exceptionally better with accuracy of 75%. Comparing the results to previous research the implemented model had performed better as the Machine Learning models had an overall score above 98%. Altough SVM takes time

Models	Accuracy	Precision	Recall	F1- Score
Random Forest Classifier	98%	99%	98%	98%
Support Vector Machine Classifier	99%	98%	99%	99%
One Class Classification (SVM)	64%	9%	25%	13%
BERT Based Functional Neural Network	75%	28%	89%	43%

Table 2: Comparison of developed model with model discussed in the literature review

but it can accurately classify the tweets and takes upper hand over Naïve Bayes which was taken into consideration by as they rejected SVM. The experiment of One Class Classification was a failure and doesn't seem to be suitable for such dataset as seen by the results it gained a F1-score of 13% for positive class.

7 Conclusion and Future Work

The research conducted thorough investigation for detection of Human Rights Violation Tweets from Twitter. This research can help NGO's and Peace Keeping Organizations to Monitor the situation where human rights are being violated. The given research was carried out using machine learning and deep learning models implemented over python.The study glimpse over the collection of data from Twitter and create a customized dataset as per the requirements. The data was further manually annotated as for the given study "Factual" Tweet were required that can be classified as "Tweet about Human Rights Violation" rather than any opinion based tweets. Various experiments were conducted where Machine Learning Model were built which took Tokenized Encoder created using TF-IDF technique whereas Deep Learning Model was built over pre-trained BERT model. From all the models Random Forest and SVM had classified data accurately where they achieved accuracy of above 98%. This accuracy was achieved as dataset was balanced using SMOTE and also Cross Validation technique of GridSearchCV was utilized which showed similar results and this shows that the model was not over-fit and was best in classification of Human Rights Violation Tweets. Also, BERT being applied over an imbalanced dataset was still commendable as model got 75% accuracy where the class weight and other parameters like optimizer and loss were considered while training the dataset. The research also showed failure of one class classification as it had the lowest F1-score and was not able to satisfy given requirements. Thus, the given research was able to achieve the objective and provided solution for the research question.

Future Work: The given research is still a prototype and has a scope of improvisation. Usage of extensive and balance dataset can be taken into consideration. The dataset can be multi-language and can be extensively collected from other Social Media like Facebook, Instagram, etc. An extended feature of verifying the text where the event mentioned is actually a fact or fake. There is scope of improvisation in the Deep Learning Models where different Deep Learning Models and Neural network can be considered

along with pre-trained models can be utilized.

Acknowledgement

I would like to appreciate and thank Dr. Rejwanul Haque and Dr. Mohammed Hasanuzaman for guidance, support, and counseling over the Research Project. Also, would like to express my gratitude to all of the activists, journalists, and especially the late Danish Siddiqui for their efforts to promote human rights.

References

Ahmad, S., Asghar, M. Z., Alotaibi, F. M. and Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques, *Human-centric computing and information sciences* **9**(1).
URL: *https://rdcu.be/ctWXk*

Alhelbawy, A., Lattimer, M., Kruschwitz, U., Fox, C. and Poesio, M. (2020). An nlp-powered human rights monitoring platform, *Expert Systems with Applications* **153**: 113365.
URL: *https://www.sciencedirect.com/science/article/pii/S0957417420301901*

Azizan, S. A. and Aziz, I. A. (2019). Terrorism detection based on sentiment analysis using machine learning, *Journal of Engineering and Applied Sciences* **12**(3): 691–698.

Bashir, S., Bano, S., Shueb, S., Gul, S., Mir, A. A., Ashraf, R., Shakeela and Noor, N. (2021). Twitter chirps for syrian people: Sentiment analysis of tweets related to syria chemical attack, *International Journal of Disaster Risk Reduction* **62**: 102397.
URL: *https://www.sciencedirect.com/science/article/pii/S2212420921003587*

Febriana, T. and Budiarto, A. (2019). Twitter dataset for hate speech and cyberbullying detection in indonesian language, *2019 International Conference on Information Management and Technology (ICIMTech)*, Vol. 1, pp. 379–382.

Fitri, V. A., Andreswari, R. and Hasibuan, M. A. (2019). Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm, *Procedia Computer Science* **161**: 765–772. The Fifth Information Systems International Conference, 23-24 July 2019, Surabaya, Indonesia.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050919318927*

Hamdan, H., Bellot, P. and Bechet, F. (2015). Lsislif: Feature extraction and label weighting for sentiment analysis in twitter, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics.

Haque, R., Hasanuzzaman, M., Ramadurai, A. and Way, A. (2019). Mining purchase intent in twitter, *Computación y sistemas* **23**(3).
URL: *https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3254*

Jiang, D., Luo, X., Xuan, J. and Xu, Z. (2017). Sentiment computing for the news event based on the social media big data, *IEEE Access* **5**: 2373–2382.

Kalliatakis, G., Ehsan, S., Fasli, M., Leonardis, A., Gall, J. and McDonald-Maier, K. D. (2017). Detection of human rights violations in images: Can convolutional neural networks help?

Kanakaraddi, S. G., Chikaraddi, A. K., Gull, K. C. and Hiremath, P. S. (2020). Comparison study of sentiment analysis of tweets using various machine learning algorithms, *2020 International Conference on Inventive Computation Technologies (ICICT)*, pp. 287–292.

Laskar, R. H. and Sunny, S. (2021). Indian journalist killed in line of duty by taliban, *The Hindustan Times* .
URL: *https://www.hindustantimes.com/india-news/indian-photojournalist-danish-siddiqui-killed-in-afghanistan-s-kandahar-province-101626420192681.html*

Lim, Y. Q., Lim, C. M., Gan, K. H. and Samsudin, N. H. (2020). Text sentiment analysis on twitter to identify positive or negative context in addressing inept regulations on social media platform, *2020 IEEE 10th Symposium on Computer Applications Industrial Electronics (ISCAIE)*, pp. 96–101.

Mandloi, L. and Patel, R. (2020). Twitter sentiments analysis using machine learninig methods, *2020 International Conference for Emerging Technology (INCET)*, pp. 1–5.

Neethu, M. S. and Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques, *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1–5.

Pandey, R., Purohit, H., Stabile, B. and Grant, A. (2018). Distributional semantics approach to detect intent in twitter conversations on sexual assaults, *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 270–277.

Patil, M. and Chavan, H. (2018). Event based sentiment analysis of twitter data, *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1050–1054.

Purohit, H., Dong, G., Shalin, V., Thirunarayan, K. and Sheth, A. (2015). Intent classification of short-text on social media, *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 222–228.

Shahare, F. F. (2017). Sentiment analysis for the news data based on the social media, *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1365–1370.

Singh, R. P., Haque, R., Hasanuzzaman, M. and Way, A. (2020). Identifying complaints from product reviews: A case study on Hindi, *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, NLP Association of India (NLPAI), Indian Institute of Technology Patna, Patna, India, pp. 108–116.
URL: *https://aclanthology.org/2020.icon-main.14*

Somula, R., Dinesh Kumar, K., Aravindharamanan, S. and Govinda, K. (2020). Twitter sentiment analysis based on us presidential election 2016, *in* S. C. Satapathy, V. Bhateja, J. R. Mohanty and S. K. Udgata (eds), *Smart Intelligent Computing and Applications*, Springer Singapore, Singapore, pp. 363–373.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, California, pp. 88–93.
URL: *https://aclanthology.org/N16-2013*

Zahoor, S. and Rohilla, R. (2020). Twitter sentiment analysis using lexical or rule based approach: A case study, *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pp. 537–542.

Zimbra, D., Abbasi, A., Zeng, D. and Chen, H. (2018). The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation, *ACM Trans. Manage. Inf. Syst.* **9**(2).
URL: *https://doi.org/10.1145/3185045*