Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 10 October 2021

Research Article

Text Classification for Human Trafficking Using Advanced Transformers

Abhinav Pandey*

Domain Expert and Author, Data Science, Machine Learning & Artificial Intelligence, India

***Corresponding Author:** Abhinav Pandey, Domain expert and Author, Data Science, Machine Learning & Artificial Intelligence, India.

E-mail: abhinav.pandey.bit@gmail.com

Received: July 15, 2021

Published: September 07, 2021

Abstract

One of the most heinous crimes of the present times, Human Trafficking has been increasing at an alarming rate globally affecting millions of men, women and children. Amongst the various types of human trafficking comprising forced employment, organ smuggling, child marriage, sex trafficking and debt bondage, the market for sex trafficking has been making the headlines worldwide. Traffickers exploit these men, women and children and force them into flesh trade sometimes as early as when a girl attains puberty. Sadly, there is truly little that any law enforcement agency could do to bring the notorious traffickers to justice as majority of the cases do not even get reported. The advent of Internet has added to the woes of law enforcement authorities as the Traffickers easily advertise online from the comfort of their homes anywhere in the world. Traffickers are easily able to dodge the authorities by continuously deploying innovative advertising patters like using non-standard English grammar, emojis, multiple victims advertised simultaneously etc. This makes it extremely difficult to filter human trafficking ads from the genuine online escort service ads. In this study, we propose a novel architecture which extends BERT to incorporate not just texts but also emojis, special characters and other advertisement language patterns for doing multi-class text classification of the online advertisements into varying possibilities of them being labelled as sex trafficking advertisements.

Keywords: Human Trafficking; Deep Learning; Natural Language Processing; Forced Labor; Emojis; Semi-Supervised Learning

Abbreviations

Ads: Advertisement; AI: Artificial Intelligence; ALBERT: A Lite BERT; AUC: Area Under the Curve; BERT: Bidirectional Encoder Representations from Transformers; CNN: Convolutional Neural Network; DFID: Department for International Development; DistilBERT: Distilled version of BERT; DT: Decision Tree; ELMo: Embeddings from Language Models; GDPR: General Data Protection Regulation; GLOTIP: Global Report on Trafficking in Persons; GPT-2: Generative Pre-trained Transformer 2; GPU: Graphics Processing Unit; HTDN: Human Trafficking Deep Network; ILO: International Labour Organization; KM: Knowledge Management; KNN: K-Nearest Neighbors; LDA: Latent Dirichlet Allocation; LM: Language Model; LR: Logistic Regression; LSTM: Long Short-Term Memory; ML: Machine Learning; NB: Naïve Bayes; NLP: Natural Language Processing; ORNN: Ordinal Regression Neural Network; RBF: Radial Basis Function; RF: Random Forest; RoBERT: Robustly Optimized BERT; ROC: Receiver Operating Characteristic; SGD: Stochastic Gradient Descent; SMOTE: Synthetic Minority Oversampling Technique; SVM: Support Vector Machine; TPU: Tensor Processing Unit; UNODC: United Nations Office on Drugs and Crime; XGBoost: eXtreme Gradient Boosting.

Introduction

One of the most heinous crimes of modern world, Human Trafficking, it is defined as use of force, of abuse of power, of deception, of abduction to gain the control or approval of a human being applying control over another human being for the objective of giving or receiving payments illegally. This crime does not spare any gender, race, or community and is rapidly growing across the globe. Please have a look into the figure 1 that shows shares of victims detected within their own country's borders, by subregion across the prominent geographies of the world.

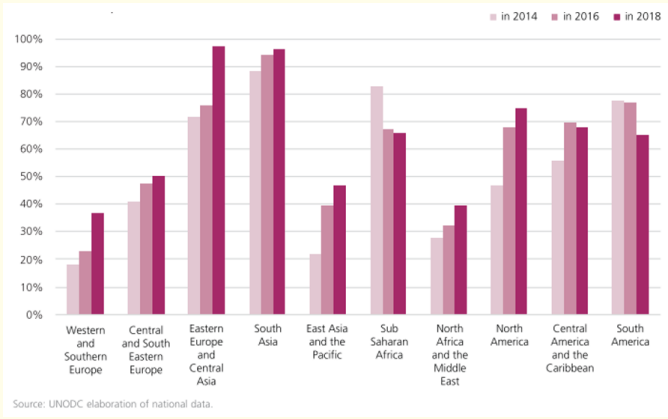

Figure 1: Victims detected within their own country [1].

This claim is supported by the fact that over the years, the conviction rate of trafficking in person has increased significantly, which is clear from figure 2 and figure 3.

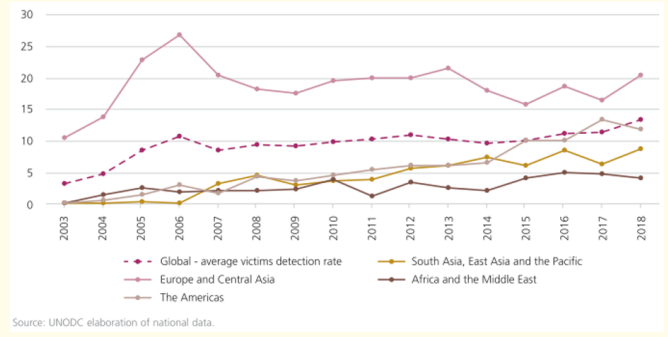

Figure 2: Detection rates for victims (per 100,000) [1].

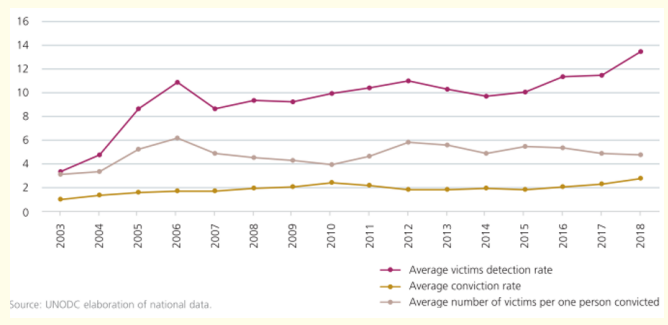

Figure 3: Victims per one person convicted of trafficking [1].

The current Covid-19 pandemic has added to the problems of millions with reports suggesting that the COVID-19 Pandemic Recession could result in a sharp decline (6.2% predicted) in GDP per capita, possibly making it the most dreaded recession since the end of WWII [1]. This suggests that individuals who are in dire need may be more likely to be ready to take risks and more likely to be exploited by the human-traffickers.

The report further states that Children have become easy target attributed to material deprivation. Amongst the different forms of trafficking that are known to exist in the world today, trafficking for flesh trade (sexual exploitation) takes the major share in the instances involving victims in economic need, when these get compared to the instances reported across all GLOTIP (Global Report On Trafficking In Persons) court cases which is manifested in the figure 4, source: [1]. The x-axis represents different forms of trafficking, and the y-axis represents the actual percentage of total cases in a particular trafficking type.

Figure 4: Exploitation forms distribution reported in court cases [1].

With the world continuously transforming digitally, the use of internet for facilitation of sex-trafficking is increasing rapidly. There is no social media platform left where the traffickers are not advertising and exploiting the victims. The internet provides the traffickers with an increased pool of sex buyers. The traffickers or pimps use internet widely to advertise their services. Every day, there are several thousand escort ads posted online. Traffickers use various keywords to disguise the trafficking ads as genuine at-will posted ads. This poses the real challenge for any law enforcement agency as to how to track down trafficking ads. One such example is the website Backpage.com which was a classified advertising website before it was shut down by federal authorities for allowing its platform to be used for illicit sexual activities. With the

limited resources, it is nearly impossible for any entity to manually scan all the text and narrow down the trafficking ads.

The motivation behind this study is to understand the indicators and other significant patterns of online sex-trafficking, the modus operandi followed by the traffickers, to understand how they use technology to continuously dodge the law enforcement authorities and continue to exploit millions globally and, to devise a system using advanced Natural Language Processing and Deep Learning methods to help law enforcement authorities to curb this illegal activity by understanding the strong indicators of online sex-trafficking in the advertisements posted on classified webpages.

Materials and Methods

In this section we shall describe the proposed method that shall be used in this study. There have been a lot of recent advancements in the field of NLP using Deep Learning mechanisms. One such model [2] was introduced in the year 2019 by scientists at Google. BERT stands for Bidirectional Encoder Representations from Transformers which was designed to pretrain deep bidirectional representations from unlabelled text using an attention mechanism that learns contextual relations between words (or sub-words) in a text. For this study, we propose to use BERT available in the Hugging Face library for the purpose of detecting online human trafficking. These models shall be used to classify the advertisements on a scale of 0 to 6 (from a "Certainly No" to "Certainly Yes"). The results obtained from these models shall be compared with traditional classification algorithms like Logistic Regression, Decision Tree, Random Forest and XGBoost (eXtreme Gradient Boosting). The following sections provide a very high-level architecture of the models for reader's reference.

Proposed extensions to BERT

In this study, we propose to extend BERT to factor in the below scenarios during encoding:

- Use of Emojis in the text corpus.
- Identifying and giving weightage to trafficking indicators like country of origin, mention of spa or massage services, use of third person language, words and phrases of interest, victim weight and advertisement language pattern.

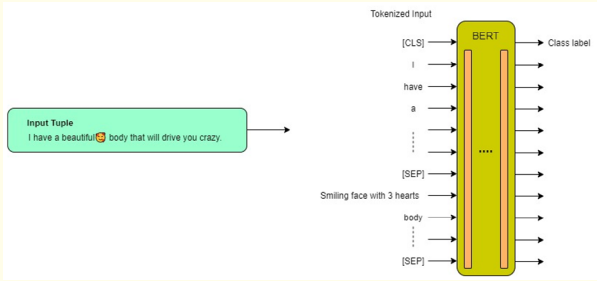

Figure 5: Proposed Architecture of our Model.

A high-level architecture of the proposed model is shown in figure 5. The input is a real advertisement from an online escort classified website. Notice that the emoji has been replaced by a descriptive token outputting the class label after a multi-class text classification process. The section below takes us through to the current stage of text classification being done on the Trafficking-10K dataset [3].

Multi-class text classification pseudo code and results

In this section, we have added the current state of our work and have attempted to explain the text classification task with the help of pseudo-code, results, and figures.

After the data cleaning is done and after splitting the dataset into train and test subsets, we apply LabelEncoding and transform the target column as shown in figure 6.

START:
INIT:
Initialize encoder variable.

COMPUTE:
Apply "fit" method on the target variable.

INIT:
Initialize encoded target (train & test) variables by applying "transform" method.

Figure 6: Pseudo code for initialization.

This results in our encoder correctly determining the classes as shown in figure 7.

```
PRINT:
Output the encoder classes.

RESULT
array([0, 1, 2, 3, 4, 5, 6])
```

Figure 7: Encoder classes.

Our current implementation is using a pre-trained BERT model "bert_multi_cased_L-12_H-768_A-12/2" which has a universal sentence encoder for 100+ languages trained with conditional masked language model to incorporate the heavy use of non-English language in the online escort advertisements.

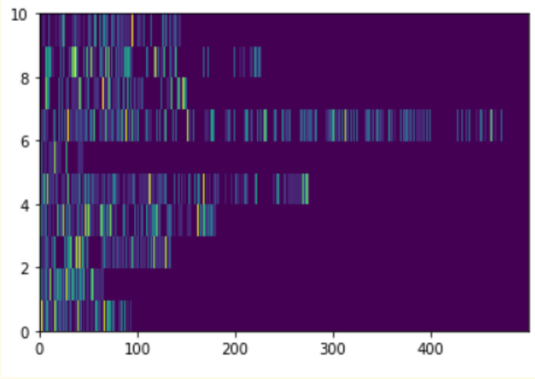

Figure 8: Tokenized advertisement text.

Once our model is loaded, we go for tokenizing the advertisement text and plot the same for a better graphical representation as shown in the figure 8. As per the process we follow in any classification task using BERT, we create the input pipeline containing input_word_ids, input_mask and input_type_ids. A pseudo-code for the same has been shown in figure 9.

```
DEFINE:
Define "inputs" variable from "input_word_ids", "input_mask" and "input_type_ids"
```

Figure 9: Defining "inputs".

This forms the base of our BERT encoding layer and then we encode the X_train and X_test to be used while we train our model. A pseudo-code (Figure 10) and a graphical representation of our model (Figure 11) have been included for better clarity.

```
INIT VARIABLES:
Initialize "num_class" as length of "encoder classes"
Initialize "max_seq_length"
Initialize "input_word_ids", "input_mask" and "input_type_ids"
from TensorFlow's Keras library

INIT OUTPUT
Initialize "pooled_output" and "sequence_output" from bert object.
Initialize "output" from TensorFlow's Keras library

INIT MODEL:
Initialize the "model" from TensorFlow's Keras library passing above
input & output layer variables
```

Figure 10: Initialize variables, output layer and model.

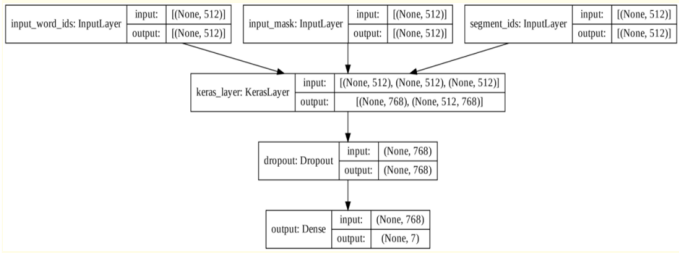

Figure 11: BERT Model Representation.

Results and Discussion

Once our model is trained, we apply the "fit" method on it to obtain the train loss and validation loss and train accuracy and validation accuracy. A pseudo-code with results has been included as shown in figure 12 for reader's reference.

```
TRAIN MODEL:

Compute "history" by applying "fit" method on our model.
Input training & validation sets with other hyperparameters like
epochs & batch size to the "fit" method.

PRINT RESULTS
Epoch 1/3
1185/1185 [==============================] - 1382s 1s/step - loss:
1.4525 - accuracy: 0.4481 - val_loss: 1.3192 - val_accuracy: 0.5089
Epoch 2/3
1185/1185 [==============================] - 1376s 1s/step - loss:
1.3019 - accuracy: 0.5079 - val_loss: 1.2733 - val_accuracy: 0.5076
Epoch 3/3
1185/1185 [==============================] - 1376s 1s/step - loss:
1.1769 - accuracy: 0.5617 - val_loss: 1.2927 - val_accuracy: 0.5139
```

Figure 12: Training the model.

The next step is for us to evaluate the model and compare the train and test accuracies as shown in the pseudo-code shown in figure 13 as well as through a graphical representation as shown in figure 14.

COMPUTE TRAINING ACCURACY:

Compute training loss and accuracy by applying "evaluate" method on our model.

PRINT TRAINING ACCURACY

COMPUTE TESTING ACCURACY

Compute testing loss and accuracy by applying "evaluate" method on our model.

PRINT TESTING ACCURACY

RESULT

Training Accuracy: 0.6118

Testing Accuracy: 0.5139

Figure 13: Computing the accuracy.

Figure 14: Plots for accuracy and losses.

After attaining the results, we go ahead and do a prediction on an actual record to see how our model is performing. Pseudo-code and steps used for prediction have been included for reader's reference. Please note that the text shown in the pseudo-code is real and contains use of explicit language, hence, reader's discretion is advised. This post does not intend to hurt sentiments of any caste, community or religion and is solely used for the objectives highlighted in this study.

On prediction, our model shows a predicted class label 2, which means the advertisement is "Weakly Not" a case of human trafficking and has likely been posted by the escort herself. The result matches with that of the actual label in the given dataset.

The results shown in figure 15 are the actual results obtained in our multi-class text classification task after applying the trained model on the chosen dataset. A comparison of the proposed model (BERT) with other traditional machine learning approaches has

TEST OUR MODEL ON UNSEEN DATA

Input an advertisement text as text = [r'Are you stressed and need Some hot reliever? I'm a horny mature girl, I love squirting I Wanna fulfill your widest sexual fantasies I'll satisfy you with my full service such as ****, oral, doggy, hardcore I can ride your till you beg me to stop., I'm honest and sincere lady you'd ever love to meet I'm available all day and night for incall and outcall... text me for rate and unforgettable experience % Real & Ready Now Naughty Fun My Place Or Yours im always available']

Initialize "inputs" variable from "bert_encode" by passing advertisement text and tokenizer as inputs.

PREDICTION

Get the prediction by applying "predict" method with "inputs" variable.

PRINT PREDICTED RESULTS

[[0.15213722 0.45104015 0.887089 0.4143432 0.7291974 0.35497358 0.10239133]]

PRINT PREDICTED CLASS LABEL

2

PRINT ENCODED CLASS

Get encoded class by applying "np.argmax" on our prediction result.

2

Figure 15: Testing our model on unseen data.

been shown in table 1. Currently, the model is getting optimized for the purpose of attaining higher accuracies by factoring in the use of emojis, advertisement patterns, countries of interest, words and phrases of interest, multiple escorts included, and language models used in the advertisement [4,5].

Conclusion

In the "Materials and Methods" section, we have outlined the complete data pipeline steps like pre-processing, transformation, treating class imbalance and finally model evaluation. We also described high-level architectures of the most recent and advanced models used in NLP for text classification tasks. We have included the current state of our text classification tasks with the help of pseudo-codes, results, and plots for reference. We finally included the proposed architecture of our model that would extend BERT to decode emojis, keeping it simple, efficient, faster, more scalable, and easily generalizable to datasets from different geographies of the world.

Acknowledgements

I must express my heartfelt gratitude to my parents and to family for providing me with unflagging support and endless encour-

agement throughout my years of study and through the process of researching and writing this thesis. In these tough times of Covid-19 pandemic, this accomplishment would not have been possible without their continued emotional support. Thank you!

Conflict of Interest

NA.

Bibliography

1. Kangaspunta Kristinia., *et al*. "UNODC global report on trafficking in persons 2020". (2021).

2. Devlin Jacob., *et al*. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding" (2021).

3. Tong Edmund., *et al*. "Combating Human Trafficking with Deep Multimodal Models". ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), vol. 1, Association for Computational Linguistics (ACL), May (2017): 1547-1556.

4. ILO. Work in Freedom Reducing Vulnerability to Trafficking of Women and Girls in South Asia and the Middle East Supporting Informed Migration, Fair Recruitment and Decent Work (2018).

5. Mccarthy Lauren A. "Human Trafficking and the New Slavery". *Annual Review of Law and Social Science* 10 (2014): 221-242.

Volume 3 Issue 10 october 2021