Sentiment Analysis On Juvenile Delinquency Using BERT Embeddings

MSc Research Project
Data Analytics

Abhinav Thapa
Student ID: 20259409

School of Computing
National College of Ireland

Supervisor: Dr. Abdul Razzaq

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Abhinav Thapa
Student ID:	20259409
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr. Abdul Razzaq
Submission Due Date:	01/02/2023
Project Title:	Sentiment Analysis On Juvenile Delinquency Using BERT Embeddings
Word Count:	6874
Page Count:	24

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:	
Date:	1st February 2023

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):	

Sentiment Analysis On Juvenile Delinquency Using BERT Embeddings

Abhinav Thapa
20259409

Abstract

Criminal Juvenile Delinquency is one of the most prevalent problems in the modern society. It generally occurs because of poverty, poor education, lack of awareness, insufficient social infrastructure, and unstable family conditions. Although there are juvenile justice systems in place, few offenders are penalized due to their young age under the guise of teenage expression for less severe crimes. However, sometimes benign crimes eventually propagate over the years into full-fledged anti-social behavior recovering from which becomes virtually impossible. Such offenders can be corrected with effective policy making. But it is difficult to draw policies balancing social conduct and human rights in case of juveniles due to which identifying public opinion in such cases is helpful in decision making. Hence, this project aims at performing sentiment analysis by fine-tuning BERT transformer models on twitter posts dataset to gauge public sentiment towards juvenile delinquency. We also propose a contrast between a BERT fine-tuned model and Machine Learning (Random Forest and Support Vector Classifier) model with BERT embeddings for sentiment classification to determine whether fine-tuning is worth the effort for this problem domain. The feature engineering approach using BERT features (100% Class Accuracy) outperformed the fine-tuned BERT model (77% Class Accuracy) on a benchmark of twitter post dataset on juvenile delinquency proving that in terms of design complexity, execution time, and performance, features engineered directly from the primary dataset using BERT has higher utility when compared with fine-tuned transfer learning approach.

1 Introduction

The world as we know it, is maintained by a delicate balance in society facilitated by justice, law, and order. Law is practiced not only to maintain the current order of things but also to promote moral values that ensure the sustenance of mankind both for present and for future generations. In societies plagued with poverty, insufficient social infrastructure, unemployment, and corrupt justice system, even a trivial event could trigger catastrophic transformations from which coming back could take decades. Hence, it is of absolute importance to not only carefully maintain the social order with proper policies but also reform them regularly. Juvenile Delinquency is one such problem in the modern world that affects almost everyone. It may not seem like a significant issue in the big picture, however, understanding it could provide valuable insights for policy making.

With around 7.8 billion total population out of which 1.3 billion (16%) juveniles in the world today, it is essential to consider and maintain proper laws and policies that affects such a group prudently. According to the World Health Organization (WHO) fact sheet on Adolescent Health 2015, approximately 430 young people aged 10-24 die every day due to interpersonal violence. Homicide is the fourth leading cause of death in people aged 10-29 years. Besides weak state policies, unemployment and socio-political causes, the general situations that push youth towards delinquency are poverty, volatile family conditions and poor education. Juvenile Delinquency is one of the most prevalent social problems in the modern society that cannot be ignored under the guise of teenage expression or aggression anymore. If left unchecked it could poison society's foundations and leave the youngsters, vulnerable.

In Ireland, some of the most common crime offences by juveniles include theft, vandalism, threat to murder, harassment, assault, and controlled drug offences. It is not that the existing policies pertaining to juveniles do not work, but the fact that the policies need to be crafted and enforced with special care and need to be updated with the current situations. Although the existing policies against juvenile delinquency do not encourage anti-social behavior, but its lack of the desired effect makes it insufficient. Because of a malleable mindset, juveniles are prone to anti-social behavior (Alatrista-Salas et al.; 2019; Akila et al.; 2020) which must be controlled by proper law enforcement without brutality. Once driven to commit a crime, they must be rehabilitated before they enter the society. Prosecuting juveniles without consideration for their potential future may lead to public uproar while leniency may lead to motivation and potential criminal behavior in the long-term. Juvenile Delinquency cannot be addressed with impunity towards the defendant whereas the judgement must deliver justice. To ensure accountability that leads to correction, striking the proper balance is important that must be maintained between severity of crime and justice delivered which may not be as easy as it seems. Simply fixing the ineffective policies may sound apparent, however, changing policies that affect such a huge portion of the society without empirical evidence and reliable scientific data could be dangerous. Hence, guiding social change while ascertaining public sentiment may provide a path to solutions for such complex problems as without proper up-to-date correction infrastructure and laws in place, it can lead to devastating impact on the future of the society we live in.

Sentiment Analysis is one of the most exciting fields for Natural Language Processing researchers (NLP) as it not only provides insights about the public sentiment and opinion about the concerned domain, but also how a certain event can be leveraged to effect changes to marketing schemes, leadership decision and even policy change. Tough decision making is imperative for any organization to thrive in a cut-throat competitive world, and even for the governments across the world it is essential to not only maintain the market stability but also manage policies to ensure growth and opportunity for each class of the society. However, it is very difficult to come up with strategies to deliver to the expectations of the citizens when the effect of the policy could change the country forever, and even more so difficult to pass it concerning juveniles. Possible solution to this conundrum would be to conduct a poll and gauge the public sentiment over a specified period and consider exceptions and mood changes and then using advanced machine learning to help aid with the solutions. BREXIT was one of popular implementations of Sentiment Analysis and its impact. Its powerful applications extend to customer experience, retail, services, leadership and, public and social decision making.

Twitter is a public forum with 1.3 billion accounts of which over 237 million daily active users, where the world population express their opinion on events in up to 280 characters. Due to the constraints in twitter posts, Twitter has become a source of unaltered data for research that could be used for extracting valuable insights about public inclinations and latest trends towards major social issues. Harnessing twitter platform for driving social change could be rewarding with regards to the problem of Juvenile Delinquency and gauging public sentiment could be a start for informed legislative policy making.

This project delivers a controlled study to bridge the gap in the existing literature about feasibility of Fine-Tuned Transformers (BERT) models (trained on publicly available datasets: Sentiment140 and IMDB) when compared with standard models (RandomForest and Support Vector Machine) ensembled with BERT Embeddings in terms of accuracy, loss, scalability, computational time and complexity.

Therefore, the purpose of this project is to make following contributions:

1. Fine-tuned BERT model on Sentiment140 and IMDB labeled datasets for text classification.

2. Framework for Feature engineered embeddings from BERT and classification using RF and SVC model.

3. Above mentioned models are assessed and contrasted for performance to determine whether fine-tuning a BERT model is worth the effort for Juvenile Delinquency when compared with model using BERT features for classification.

4. Benchmarked Twitter dataset annotated with sentiment classes: positive, neutral, and negative.

The rest of the project is summarized in the following sections: Section 2: Systematic review of past literature on Sentiment Analysis; Section 3: Research methodology and experimental setup; Section 4: Evaluation Results; and Section 5: Discussion; and 6: Conclusion and future work.

2 Related Work

Sentiment analysis is one of the most popular applications of natural language processing (NLP) in which an algorithm accepts text as input and provides its inherent sentiment class as the output. It has proven to provide exceptional utility in various domains such as quality management, customer service, market research, brand management, customer feedback, infectious disease sentiment, financial sentiment, etc. It also has the capacity to gauge customer opinion about the organization's actions and provide valuable insights for informed decision making. In NLP, the goal of a model for sentiment analysis is to represent and identify the meaning of both words as well as the context of the input text to calculate the polarity of sentiment inherent to it. Sentiment analysis can be performed by various approaches such as Supervised Learning, Unsupervised Learning, Lexicon based learning, Deep Learning and Transfer learning.

Existing machine learning models such as Decision trees (DT), Naïve-Bayes (NB), Stacked classifier, Logistic Regression (Mohammad et al.; 2013; Shamrat et al.; 2021; Bouazizi and Ohtsuki; 2019; Naseem et al.; 2021; Phu et al.; 2017; Singh and Tripathi; 2021; Balabantaray et al.; 2012; Hossain and Rahman; 2022; Alatrista-Salas et al.; 2019), deep learning models such as Convolutional Neural Network (CNN), Recurrent Neural Networks (RNNs) (Tang et al.; 2015; Glorot et al.; 2011; Tang et al.; 2014; Ombabi et al.; 2020; Kumar et al.; 2022; Vaswani et al.; 2017; Monika et al.; 2019), and VADER (Valence Aware Dictionary and Sentiment Reasoner) and TextBlob lexicon and rule-based libraries are individually competent at text classification, however, large amount of labelled textual data is required to create an effective and reliable model for desired output. Since manual annotations are highly expensive, models trained sufficiently on quality corpus are essential for effective Sentiment Analysis which must be both computationally feasible and accurate. Hence, in this research we delve into a comparison between a fine-tuned BERT model and a model using BERT embeddings in RandomForest for classification to gauge the feasibility in terms of computational expense, execution time and accuracy scores for results.

This project delivers a controlled study to bridge the gap in the existing literature about feasibility of Fine-Tuned Transformers (BERT) models (trained on publicly available datasets: Sentiment140 and IMDB) when compared with standard models (RandomForest and Support Vector Machine) ensembled with BERT Embeddings in terms of accuracy, loss, scalability, computational time and complexity.

In this section, we focus on the various approaches to sentiment analysis and provide an analysis of the existing techniques and evaluate our proposed method.

2.1 Supervised Learning

In supervised learning, the model is learned on a sufficient size of training records labelled manually and tested on a validation or test set using machine learning algorithms for sentiment analysis. Some of the most popular supervised learning algorithms for sentiment analysis are as follows:

2.1.1 Tree-based Learners

These methods utilize tree-like approach to classify texts using information gain on each word representations. (Phu et al.; 2017) used Iterative Dichotomiser-3 algorithm for sentiment prediction of English documents that outputs positive, negative, and neutral sentiment classes. (Bouazizi and Ohtsuki; 2019) used randomforest for multi-class classification of Twitter data and analysed the effect of multiple class labels (Love, Hate, Happiness, Sadness, Fun, Anger and Neutral) on the classification performance. Randomforest and Decision Trees (Mohammad et al.; 2013) algorithms were used for evaluation of Sentiment classifiers on Twitter dataset. Although, such models are fast and computationally inexpensive their performance for NLP tasks is limited by the availability of labelled dataset.

2.1.2 Rule-based Learners

A set of predefined rules employing if-else based approach is followed for text classification using words related to the ruled lexicons. (Mohammad et al.; 2013) devised a lexicon based on customer feedback and used it for sentiment classification. Nevertheless, results were improved however, the approach was highly lexicon dependant and lacked scalability to other domains.

2.1.3 Other ML (Machine Learning) models (SVM, NB, LR, KNN)

SVM classifies by passing a hyperplane through a spatial representation of the text for classification. (Mohammad et al.; 2013) used SVM for sentiment classification by representing text representations into lexicon n-grams features. (Hossain and Rahman;

2022) used SVM for multi-class emotion mining from Twitter dataset. (Singh and Tripathi; 2021) employed the SVM on Term Frequency- Inverse Document Frequency (TF-IDF) cleaned McDonald's and KFC customers dataset for achieving comparative results. (Hossain and Rahman; 2022) demonstrated the comparison among machine learning algorithms, of which logistic regression performed the best on the yelp financial review dataset annotated by AFINN and VADER methods. (Shamrat et al.; 2021) performed sentiment analysis using K-Nearest Neighbour for determining the public sentiment towards covid vaccines namely Pfizer, Astra Zeneca and Moderna whereas [(Naseem et al.; 2021) created a benchmark dataset called COVIDSenti that had manually annotated 90000 tweets related to covid-19 and evaluated various state-of-the-art models.

2.1.4 Deep Learning (ANNs, RNNs)

Artificial Neural Networks (ANN) inspired by the workings of human neurons utilizes word encodings as tensors (matrix) for classification tasks. (Tang et al.; 2015) utilized gated recurrent neural network (GRU) on Yelp and IMDB movie datasets for review sentiment classification at the document-level. (Glorot et al.; 2011) used the autoencoders for classification. (Tang et al.; 2014) trained their model on ten million twitter posts to extract sentiment specific word embeddings for binary classification, while, (Ombabi et al.; 2020) utilized the CNN and LSTM features using SVM classifier for sentiment analysis. (Umer et al.; 2021) utilized the CNN features for sentiment prediction using Long Short-Term Memory Networks and provided a comparative analysis with experimentations with TF-IDF processed dataset and concluded that CNN used with other sequential models (RNN, GRU, LSTMs) outperforms most standard Machine Learning models (Logistic Regression, Support Vector Machine, RandomForest, KNN) and standard deep learning models (Multi-Layer Perceptron, LSTM, GRU). Even though the sequential learners (RNN, LSTM, GRU) (Monika et al.; 2019; Sinha et al.; 2022) can learn the contextual data better, its capacity is limited in its architecture due to which it fails to learn the long-term dependencies in a medium-large text input as the input passed needs to maintain its sequence for processing.

2.2 Unsupervised learning

2.2.1 Sentiment Lexicon based method

In this approach, a lexicon i.e., a collection of words with polarities annotated depending on specific domains is used for calculating the input words associations to the words in the lexicon to assign the suitable polarities. These can be created either manually, using dictionary method or based on language corpus. In the manual approach, the compiler manually annotates the words in the lexicon for analysis. In a dictionary approach, the lexicon is prepared by starting with a set of words with polarities and extends by associations for appropriate meaning from other dictionaries to gauge the sentiment polarities of the supplied text, while the corpus-based approach simply uses an arbitrary sample of the words in the language for analysis. (Mohd et al.; 2022) uses semantic extensions of lexicons for feature engineering to create 'lexico-semantic' features for classification. They propose a domain-independent semantic-model-based approach that assumes that words used for a certain context have similar meanings in language, to improve the accuracy of the classification models. Extensive experiments on cross-domain datasets validated the approach and outperformed the existing machine learning models. (Asghar et al.;

2017) introduced the lexicon-enhanced rule-based framework with domain specific terms including emoticons, negations and modifiers to overcome the challenges presented by sparseness of the predictions made by unsupervised learning techniques. Sentence and review level sentiment classification of review was done using the SentiWordNet-based classifier. However, it is limited by the domain specific sentiments annotations made manually which is expensive and time consuming. (Mukhtar et al.; 2018) built a wide expansive Urdu lexicon for sentiment which was used to develop the Urdu Sentiment Analyzer with a corpus of over 6000 sentences collected and annotated manually. This sentiment analyzer outperformed the supervised machine learning algorithms (DT, KNN, SVM). (Marcec and Likic; 2022) used the AFINN lexicon-based approach to study the public sentiment towards covid vaccines for the duration of four months. This approach even though increased performance, their application is restricted to the specified domains for which training was done.

2.2.2 Clustering Machine learning

It is an unsupervised learning method in which the clusters of similar entities are formed based on their closeness of relationships among them without clearly defined labels. The polarities are assigned with respect to the corresponding similarity of the input text with the cluster centroids (positive or negative). (Suresh et al.; 2016) performed a sentiment classification by applying fuzzy clustering technique whereas (Hassan et al.; 2016) utilized a density-based clustering model for text classification. Although clustering does not need labelled datasets, they are ineffective for sparse data and lack sophistication than other language models such as Transformer based models.

2.3 Unsupervised feature-based Learning (GloVe, Word2Vec, ELMO)

The competence of each of the mentioned models rely essentially on the learned vector representations of the training textual data for classification. These vectors are often called Embeddings. Since the capacity of models for learning textual patterns is as good as the vectors representing the data, using sophisticated pre-trained vector representations (GloVe, Word2Vec, ElMo) for detecting sentiment is better than learning from scratch. These text embeddings can represent not only the meaning but are also adept at encoding the contextual patterns. (Naseem et al.; 2019; Naseem and Musial; 2019) concluded in their research that with the help of such powerful feature-based text embeddings the performance for text analytics could be boosted to deliver desired results when compared with traditional lexicon approach.

GloVe (Global Vectors for Word Representations), Word2Vec and ELMo are a few popular models that learn not only the syntax but also the semantics of the text to provide embedding features that can deliver exceptional results for various NLP tasks when used in conjunction with other NLP models. RNNs (Monika et al.; 2019), even though better at extracting sequential context when compared with Bag of Words (BOW) approach, fail for medium-long text inputs and are computationally expensive. Whereas transformer models based on attention learning (Vaswani et al.; 2017) overcome this limitation by learning through parallel processing to deliver excellent performance with reduced computational time.

2.4 Ensemble Learning

In Ensemble learning approach the models are designed to learn from multiple weak learners (models) and perform classification by integrating different learning methods to provide improved results. Typically, it is used to capitalize on the expertise of different approaches to produce a combined model (strong learner) that is better than the individual (weaker) learners. (Al-Hashedi et al.; 2022) used the Word2Vec with NaiveBayes for sentiment classification on an oversampled Arabic tweets dataset and achieved significantly improved accuracy compared to individual models. (Umer et al.; 2021) used CNN with LSTM, (CHIHAB et al.; 2022) examined implementation of Bi-directional LSTM (Long Shot-Term Memory) with Multiple Linear Regression for improving on contextual feature extraction, (Kumar et al.; 2022) evaluated the LSTM and GRU (Gated Recurrent Unit) neural networks on twitter sentiment classification, (Kazmaier and van Vuuren; 2022) analysed combinations of simple voting, weighted voting, and meta-learning ensemble models while (Kazmaier and van Vuuren; 2022) evaluated Bagging, Boosting and Random Subspace ensemble methods on Decision Tree, KNN and SVM for classification. (Subba and Kumari; 2022) implemented a heterogeneous stacking of LSTM, GRU and Bi-GRU as base learner with an LSTM meta learner on an integrated word embedding of GloVe, Word2Vec and BERT for classification. Ensemble Learning is highly effective at boosting performance; however, design complexity and computational expense can increase very quickly for such models due to additional processing with each layer of added learners.

2.5 Transfer Learning

When a certain machine learning problem involves learning from a vast pool of training data to produce reliable results, it can get highly expensive in terms of computational resources and training time very quickly, due to which transfer learning has become one of the most popular training techniques for researchers in the field of NLP. Transfer learning allows to utilize the learned knowledge from one task and apply it to another related task. It not only allows model reusability, but also placates the need for expensive data collection in data-sparse domains of study.

The state-of-the-art BERT (Bidirectional Encoder Representations from Transformers) based on attention mechanism (Vaswani et al.; 2017), unlike earlier Recurrent Neural Network models (LSTM, Bi-LSTM, CNN-LSTM, GRU) specialized for sequence (context) learning in both directions, can utilize the computational power of the Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs) for parallel processing to produce reliable results faster maintaining both the contextual knowledge and vocabulary for long-term dependencies. BERT models are pre-trained on unannotated text data using Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) algorithms for customized embeddings which can then be fine-tuned for specific NLP tasks. (Sahar et al.; 2022) used transfer learning for sentiment classification on Amazon product reviews using BERT with 93.2% accuracy, (Bansal et al.; 2022) highlighted that the domain-specific BERT models (CT-BERT, BERTweet) outperform the general-purpose BERT models such as RoBERTa, BERT and XL-Net transformer models, and overcame the imbalance in dataset for classification, (AlBadani et al.; 2022) illustrated the application of fine-tuned Universal Language Model (ULMFit) based on AWD-LSTM and multiple ensembles of RNNs with SVM for document level sentiment detection on Twitter US Airlines, IMDB, and GOP debate datasets, (Chiorrini et al.; 2021; Kodiyala and Mercer; 2021) implemented BERT fine-tuning for emotion and sentiment detection with data augmentation while (Alammary; 2022; Prottasha et al.; 2022) extrapolated the BERT models implementations to Arabic and Bangla language modelling for sentiment classification. In NLP, models need to be trained with huge volumes of data to achieve reliable predictions, however, with BERT, training with even fairly smaller datasets could lead to reliable outcomes on even data-sparse domains of study such as Juvenile Delinquency.

In this research, we demonstrate the use state-of-the-art BERT model for fine-tuning on Sentiment-140 Dataset and compare performance with a model that uses BERT embeddings for classification using RandomForest on a Twitter Dataset, to determine the feasibility of the Fine-tuned model and ascertain whether its reasonable to pursue fine-tuning over using BERT embeddings for classification on Juvenile Delinquency.

3 Methodology and Design Specifications

This section describes the methodology adopted for implementation of project objectives to perform Sentiment analysis on Juvenile Delinquency and study the feasibility of Fine-tuning approach over Feature Engineering using Bert embeddings approach for Sentiment Analysis. The following model architecture inspired by Knowledge Discovery in Databases (KDD) methodology was adopted and replicated for experimentation (Figure 1).

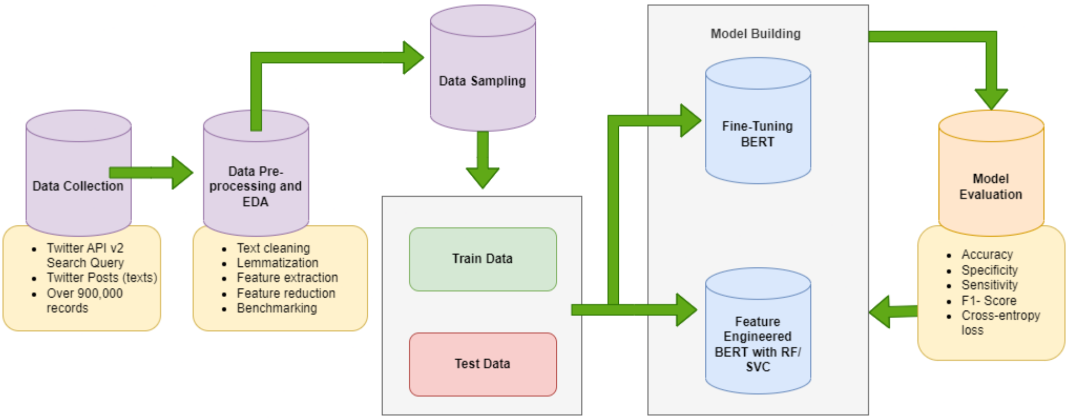


Figure 1: Project Methodology Framework

3.1 Data Collection and Preparation

The primary dataset chosen for this study was extracted from an API called "Twitter API v2", provided by Twitter for research communities. Twitter puts strict limits on scraping data without necessary permissions and bans the user if the permissions are misused. Access to their database is controlled under three access levels depending on the user application request: Essential, Elevated and Academic Research. We were approved for Academic Research license, according to which we received access to 10 million tweets per month per project.

Data collection using Twitter API v2 was done using query builder feature of the Tweet Downloader endpoint by passing customized search query for our specific problem domain.

	Hastags	Keywords
Twitter API v2 Query Builder search parameters	#Juvenilecrime #Juvenile Delinquency #Juvenile Justice #Crimehasnoage #Juvenile #Youthjustice #youthcrime #Juvenileoffenders #youngoffenders #Juvenilecourt #Youth diversion	Vandalism, Theft, Drugs, Robbery, Burglary, Attack, Assault, Damages, Violence, Juvenile, Sex offence, Homicides, etc.

Table 1: Twitter API Search Parameters

3.1.1 Search Query Design

Tweets were filtered using specific hashtags and related keywords that were carefully selected post-analysis of random hashtag search on twitter, Central Statistics Office (CSO) and RTE news. Hashtags such as Juvenilecrime, Crimehasnoage, youthcrime, youngoffenders, etc. along with keywords matching juvenile, vandalism, drugs, robbery, theft, murder, etc. were used in search query to extract related tweets from 29 May 2022 to 20 Nov 2022 and were saved as a JSON file. Over 900,000 records were extracted describing id, author id, created date, language, author data, geolocation data as features besides 'text' data, as a result of the search query implementation (Table 1).

3.1.2 Data Pre-processing

Extracted raw twitter data was cleaned, filtered, and prepared for further analysis using various natural language processing techniques and following steps were taken:

a. Although filter for fetching tweets in English language was already applied in search query, tweets with multiple languages can seep into the extracted dataset. Hence, the tweets were filtered by 'lang' column manually and posts with less than three word-count were removed.

b. Duplicate tweets were removed along with the rest of the columns except text from twitter posts.

c. Retweets starting with 'RT' were removed to eliminate further redundancy.

d. URLs, mentions, hashtags, emojis, smileys, digits, stop words and reserved words ('FAV') were removed from twitter posts.

e. Text were converted into lower case and punctuations removed.

f. Extended words were shortened (e.g. 'Helloooooo' to 'Hello'), shorthand texts were rephrased (e.g. '2moro' to 'tomorrow'), text tokens (nltk tokens, lemmatized tokens etc.) were generated and the final pre-processed dataset was stored.

3.1.3 Benchmarking Primary dataset with Labels and EDA

Since the extracted twitter dataset was unlabelled, we used the VADER (Valence-Aware Dictionary and Sentiment Reasoner) and RoBERTa (Robustly Optimized Bidirectional Encoder Representations from Transformers) models to label the pre-processed dataset with Sentiment classes (0 – Negative and 1 - Positive). This dataset was saved and used as a benchmark for evaluating model performances in the subsequent sections.

Dataset	Samples	Features	Target Labels	Class Distribution	Source
Twitter (Raw)	9,32,778	9	Unlabeled	NA	Twitter API
Twitter Benchmarked	1,63,626	2	[0- Negative,1- Positive]	84%- Negative, 16%- Positive	Processed manually
Sentiment140 (Downsized)	92,325	3	[0- Negative,1- Positive]	50%- Negative, 50%- Positive	Kaggle repository
IMDB Dataset	40,000	2	[0- Negative,1- Positive]	50%- Negative, 50%- Positive	Kaggle repository

Table 2: Project Data Description

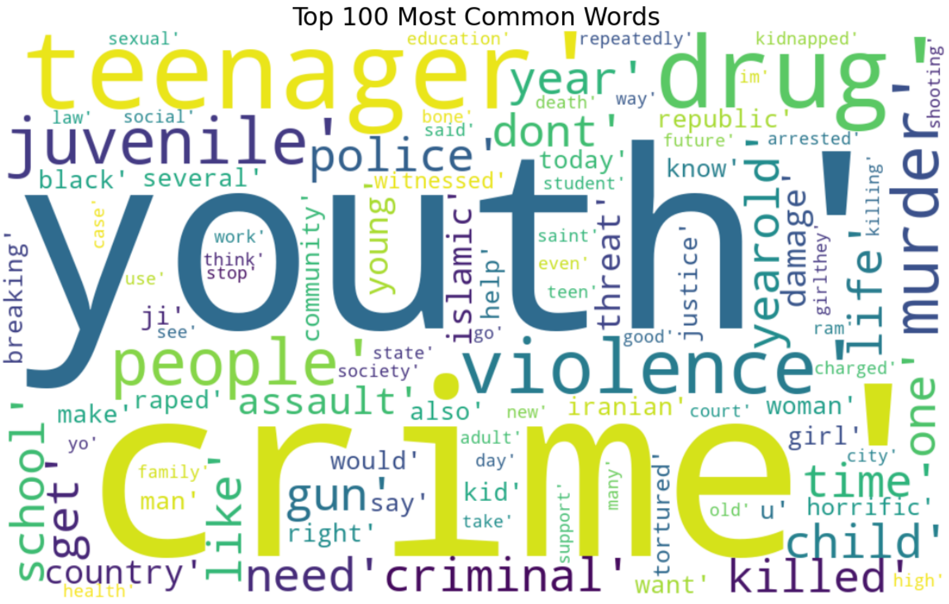

Top 100 Most Common Words

Figure 2: Word Cloud for Twitter Benchmark

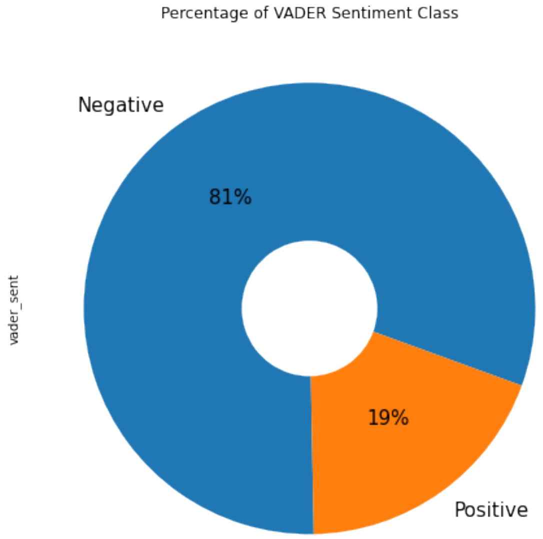

Percentage of VADER Sentiment Class

Figure 3: Class Distribution : VADER

RoBERTa Scores were used to label the twitter dataset for benchmark purposes for evaluation instead of VADER scores, due to the limited capacity of lexicon-based models when compared with transformer-based models to extract contextual meaning from

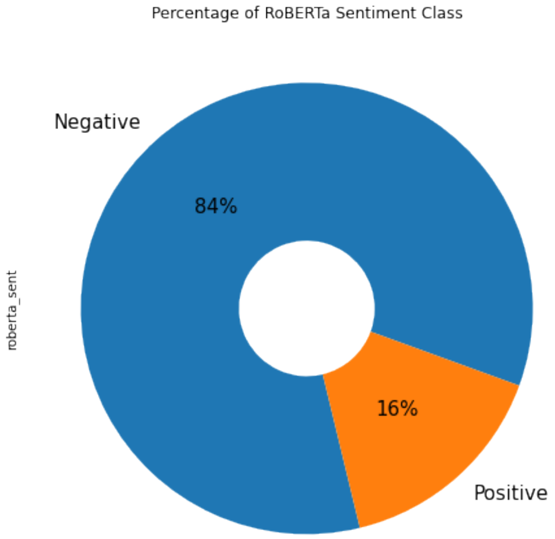


Figure 4: Class Distribution : RoBERTa

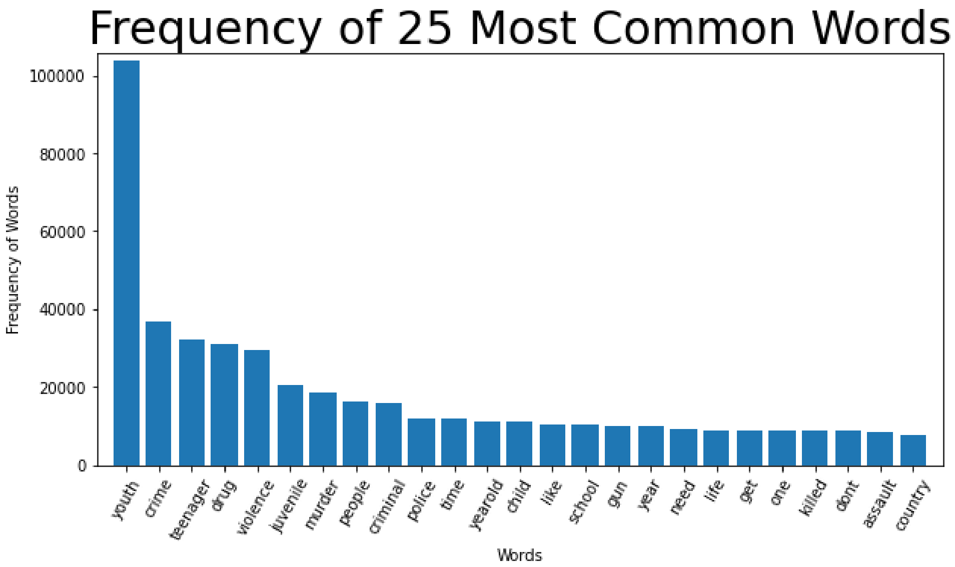


Figure 5: Common Features

the input text. Apart from Benchmark Twitter data, publicly available labelled data Sentiment-140 and IMDB data were used as secondary data source for model training. Following features were observed (Table 2, Figure 2-5).

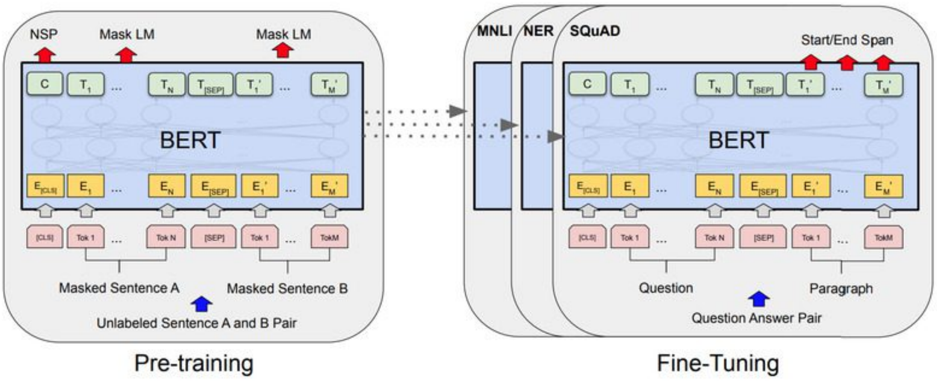


Figure 6: Google BERT Model Architecture

Parameter	Count
Transformer Layers	12
Hidden Nodes Size	768
Self-Attention Heads	12
Total Parameters	110 million

Table 3: Pre-Trained BERT Base Uncased Model Parameters

3.2 Model Building

3.2.1 Fine-Tuning BERT with Sentiment140 (Approach 1)

BERT (Figure 6[1]) models are different from other ML models in that they can utilize the concept of transfer learning exceptionally well and can be trained (fine-tuned) to perform significant results for NLP tasks. In this approach, we fine-tuned custom BERT ('bert-base-uncased') model with Sentiment140 dataset[2] for predicting class labels on the twitter benchmark dataset. The proposed custom BERT model inherited parameters from huggingface 'bert-base-uncased' model and compiled several customized layers including a dropout layer, a rectified linear (Relu activation) layer, a fully connected layer (size 768x512), another fully connected layer (size 512x2) and a softmax (activation) layer for conversion. This custom model architecture was initialized with AdamW optimizer for classification.

Due to the computational limitation of the local system and cloud GPUs in this study, we downsized the Sentiment140 dataset from having 1.6 million records to 100,000 records and pre-processed for training on BERT. The pre-processing included similar steps used

[1] Refer : https://www.kdnuggets.com

[2] https://www.kaggle.com/

Parameter	Count
Fully Connected Layer	762
Activation	ReLu
Fully Connected Layer	512
Output (Activation)	2 (Softmax)
Optimizer	AdamW

Table 4: Custom Fine-Tuned BERT Model Parameters

for pre-processing Benchmark Twitter dataset except that the downsized Sentiment140 dataset was transformed to have a nearly balanced class (1:1 for Positive and Negative) distribution.

The pre-processed Sentiment140 dataset was split into train, validation, and test sets for model training, tokenized using BERT tokenizers, and converted to tensors for model training. Customized methods for model training on the GPU was formulated and executed for 10 epochs and the results were obtained. Final models with least loss were saved for re-usability.

Similar experiments were replicated using IMDB dataset[3] for fine-tuning BERT base uncased model for evaluation on Twitter benchmark dataset and the results were recorded for evaluation.

For Sentiment140 tuned model predictions on benchmark dataset, the dataset was downsized owing to the computational capacity of the available infrastructure and to save on execution time. Saved Fine-Tuned BERT model was loaded, and predictions were captured for the benchmark dataset.

3.2.2 BERT Embeddings for ML classification using RF/ SVC (Approach 2) on Twitter Benchmark Dataset

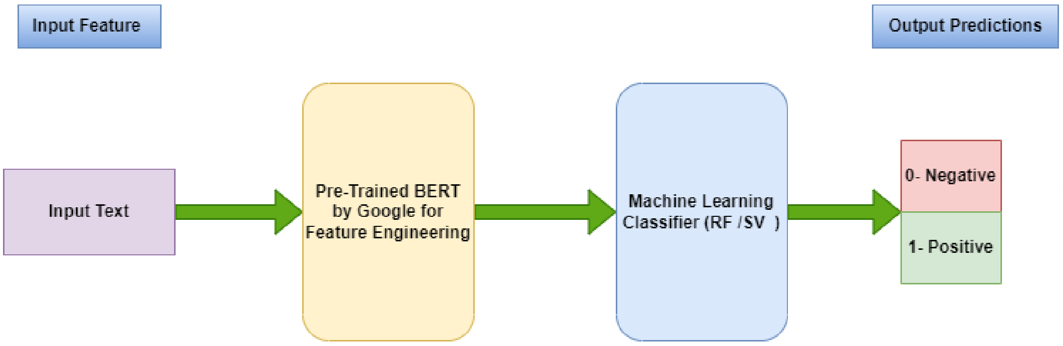

Input Feature
Output Predictions
Input Text
Pre-Trained BERT by Google for Feature Engineering
Machine Learning Classifier (RF /SV)
0- Negative
1- Positive

Figure 7: Custom RF/SVM Classifier architecture with BERT embeddings

In this approach (Figure 7 [4]), the BERT Embeddings were used for feature extraction directly from the downsized twitter benchmark dataset (14000 records) and used as input for the RandomForest and Support Vector Machine Classifier models.

Input features (text) from Twitter benchmark dataset were feature engineered using BERT uncased model tokenizers, and encoders, and saved for reuse. These tensors of training features were then flattened and passed to the classification models for training and evaluation. Trained models were saved, and predictions were made on the twitter benchmark dataset.

3.3 Evaluation Metrics

1. Accuracy: Classification Accuracy can be defined as the capacity of a classifier model to correctly predict a target label. It is one of the most widely used metric for

[3] https://www.kaggle.com/

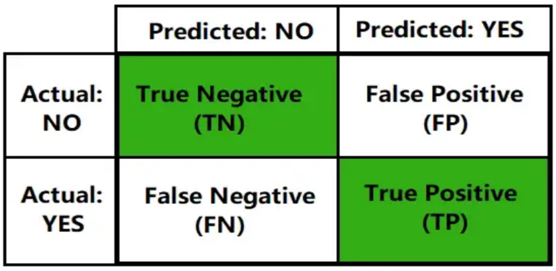
[4] https://jalammar.github.io/illustrated-transformer/

model evaluation on balanced datasets. Mathematically, it is given by the below formula:

$$Accuracy = \frac{Total\ number\ of\ correctly\ predicted\ samples}{Total\ number\ of\ samples}$$

Figure 8: Accuracy Score

2. Confusion matrix: This metric provides the complete distribution of predictions in a tabular format that can be used to calculate metrics such as accuracy, precision, recall and F1 score. It also helps detect Type I and Type II Errors in classification problems.



	Predicted: NO	Predicted: YES
Actual: NO	True Negative (TN)	False Positive (FP)
Actual: YES	False Negative (FN)	True Positive (TP)

Figure 9: Confusion Matrix

3. Precision (Specificity): This metric helps to determine the model competence for predictions of True Positives i.e. actual positive class predicted correctly. Its given as:

$$Precision = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Positives\ (FP)}$$

Figure 10: Precision Score

4. Recall: Also known as Sensitivity, it is defined as the capacity of model to predict maximum True Positive classes correctly.

$$Recall = \frac{True\ Positives\ (TP)}{True\ Positives\ (TP) + False\ Negatives\ (FN)}$$

Figure 11: Recall Score

5. F1 Score: It is the harmonized mean recall and precision metrics that ranges between 0 and 1. It can be used as a single metric for model classification competence like accuracy score for model evaluation. It is defined as:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Figure 12: F1 Score

6. Cross Entropy Loss: It is the classification metric used to calculate log loss for model output probabilities lying between 0 and 1. For binary classification it is defined as:

$$-(y\log(p) + (1-y)\log(1-p))$$

Figure 13: Cross-Entropy Loss

3.4 Experimental Design and Implementation

A similar approach was employed with IMDB dataset for fine-tuning BERT model for multiple experiments on the Twitter dataset using combinations of parameters for an elaborate model building phase. The fine-tuned models were contrasted against the machine models (RF and SVM) using BERT embeddings as features engineered for text classification. Model performances were evaluated on the Twitter Benchmarked Dataset and the results were recorded.

The proposed framework in this research involved processing of huge volumes of input data and corresponding feature representations that could not be handled with the local infrastructure (CPU Intel Core i5 8 Gigabytes), and hence cloud infrastructure (Google Colab Pro+) was employed for design and implementation. Due to the limited availability of high performing cloud processing units, the research was downsized to a controlled study for achieving the proposed objectives. It was also noted that the execution runtime may vary drastically in case the employed cloud infrastructure was not available.

Even though the advanced and powerful GPUs from Google Colab (A100-SXM4-40GB, NVIDIA-SMI 40.32.03, CUDA Version 11.2; RAM 89.6 Gigabytes) reduced a lot of training and processing time, frequent system crashes and runtime availability issues were observed throughout the research. A total of 500 compute units provided by Google Colab for a Pro+ membership was utilized for this research to achieve best results.

The study was conducted using following tools and library packages for implementation:

3.4.1 Python Language

The project was developed using Python programming language in Jupyter Notebook and Google Colab Notebook. Initial stages were developed in local notebooks, however, owing to the increased complexity during model training and evaluation stages, the project was migrated to a cloud setup and developed using Google Colab Pro+ Notebooks.

3.4.2 Essential Libraries and Packages

Since complex NLP tasks are handled exceptionally well by tensor libraries, the study primarily utilized Pytorch for model building and evaluation. Several necessary libraries explored and used were as follows:

a. Pytorch: This library was used to handle huge volumes of matrix (tensors) operations throughout the study. Neural Network module from torch library was used for custom fine-tuning BERT models. Also, tensor operations were handled in batches using Pytorch Dataloaders.

b. Tweepy: Library was used for data collection from Twitter API v2. The API user interface for query building and extraction were also used and tested using Postman API testing tool.

c. Transformers: Transformer libraries were referenced and evaluated on hugging-face repositories for this research. Several models such as BERT, roBERTa and DistilBERT were evaluated, and few chosen for the purposes of this research.

d. Integrated Development Environments (IDE): Anaconda Jupyter Notebook and Google Colab Pro+ environments were utilized for project development.

e: Pandas, Numpy, Scikit-Learn, Matplotlib, Seaborn: Common libraries for data pre-processing, transformation, visualizations, and modelling were also implemented.

f. NLTK: Natural Language Toolkit library was used for text pre-processing (text cleaning, tokenization, lemmatization, etc.), visualization (word clouds, topic modelling, etc.) and benchmarking twitter dataset.

g. VADER: Lexicon based sentiment models from NLTK library were also explored in this project.

4 Evaluation

4.1 Experiment 1: BERT model Fine-Tuned on Sentiment140 vs BERT Embedded Randomforest and SVC

Here we compared performances from a BERT model fine-tuned (approach1) on Sentiment140 dataset for five and ten epochs respectively with the (approach 2) model using BERT embeddings with Randomforest (RF) and Support Vector Classifier (SVC) respectively. Following results were observed:

4.2 Experiment 2: BERT model Fine-Tuned on IMDB dataset vs BERT Embedded Randomforest and SVC

Above setup was replicated for fine-tuning BERT (Approach 1) with IMDB dataset for evaluation with the Approach 2 models (BERT embeddings with RF and SVM) and the following results were obtained.

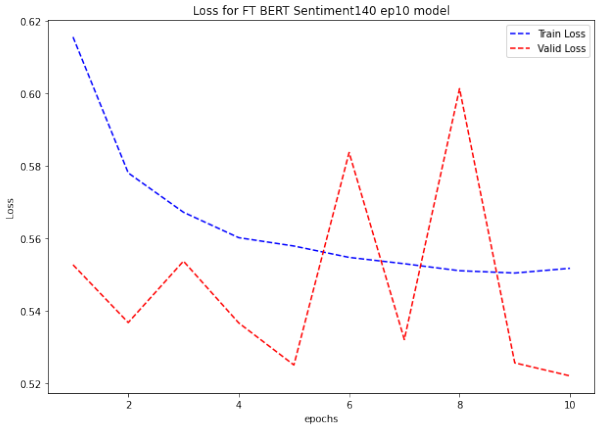


Figure 14: Loss for Fine-tuned BERT (Sentiment140) for 10 epochs

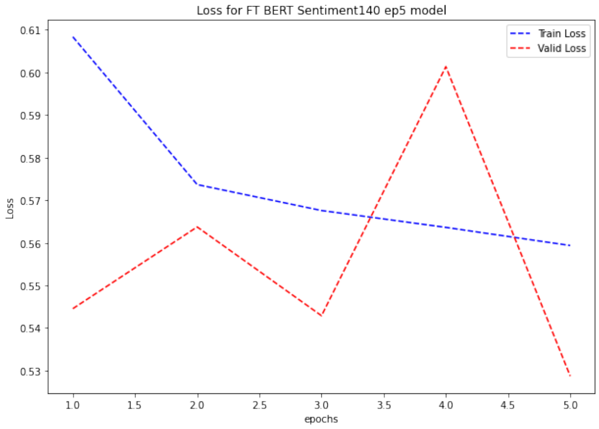


Figure 15: Loss for Fine-tuned BERT (Sentiment140) for 5 epochs

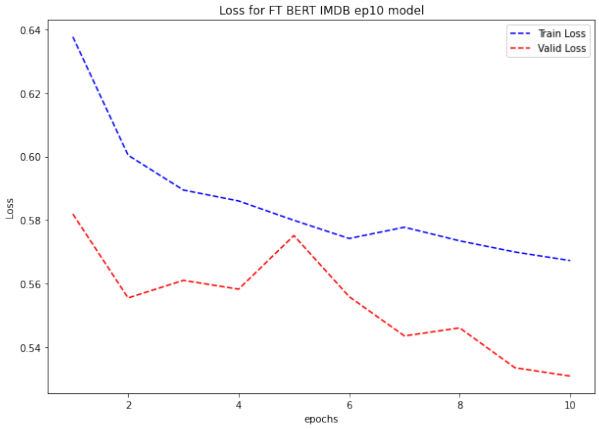


Figure 16: Loss for Fine-tuned BERT (IMDB) for 10 epochs

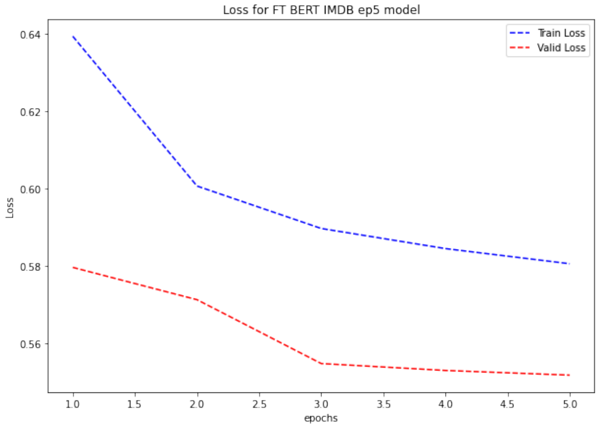


Figure 17: Loss for Fine-tuned BERT (IMDB) for 5 epochs

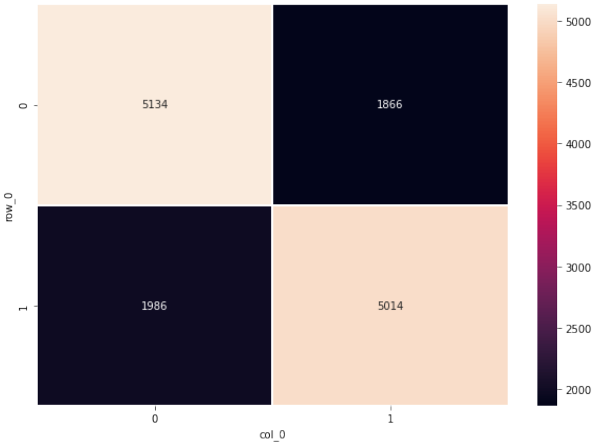


Figure 18: Confusion Matrix: Fine-tuned BERT (Sentiment140) for 10 epochs on Benchmark data

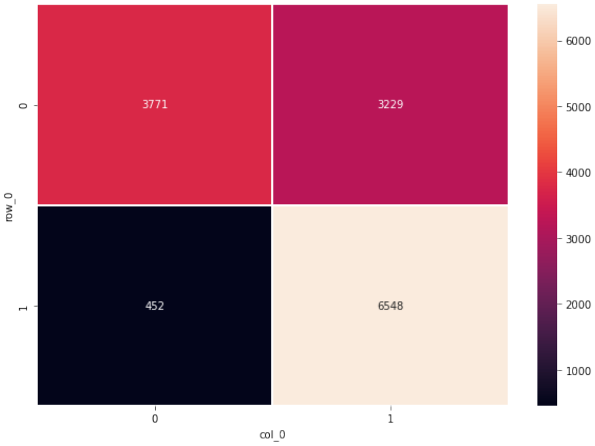


Figure 19: Confusion Matrix: Fine-tuned BERT (IMDB) for 10 epochs on Benchmark data

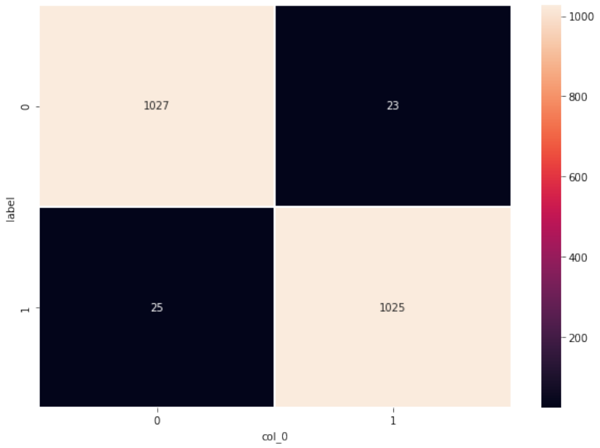


Figure 20: Confusion Matrix: RF-BERT on Benchmark data

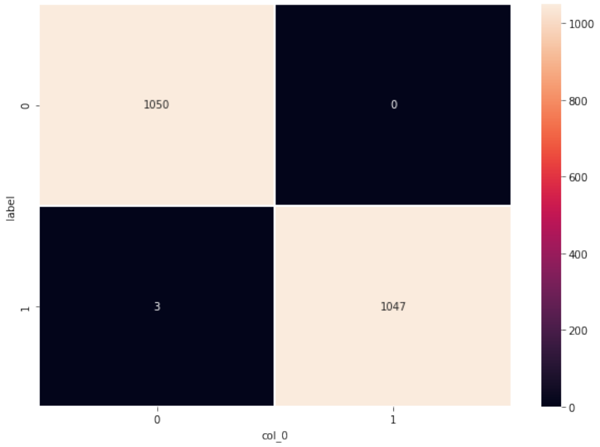


Figure 21: Confusion Matrix: SVC-BERT on Benchmark data

4.3 Discussion

Implementation and evaluation of selective experimental models were managed using available computational infrastructure to perform sentiment analysis on juvenile delinquency and results were evaluated to reveal that, although the overall classification accuracy of the fine-tuned models (77 %, highest for Model 7 and lowest 70% for Model 6) increased with the increase in number of training epochs, its performance lagged compared to the standard machine learning classifiers (RF- 98% and SVC- 100%) employed with BERT embedding features. Not only did the accuracy performance suffer but the precision and recall scores were not satisfactory compared to the RF and SVC models (Models 9-10 in Table 5,6). It was also observed that increase training epochs and parameter tuning for fine-tuned models indeed improved overall performance, it also contributed to higher model complexity, exponential computational time, and resource utilization. However, performing sentiment analysis using BERT embeddings as features extracted directly from the concerned training data and then feeding it to the standard classification models (RF, SVC, NB, KNN) yielded much better results and provided faster conversion time.

Metrics/ Models	Overall Accuracy	Class Precision (Neg/Pos)	Class Recall (Neg/Pos)	Class F1-Score (Neg/Pos)
Model 1	0.73	0.72 / 0.76	0.79 / 0.68	0.75 / 0.72
Model 2	0.72	0.72 / 0.73	0.73 / 0.72	0.73 / 0.72
Model 3	0.74	0.73 / 0.76	0.78 / 0.71	0.76 / 0.73
Model 4	0.73	0.73 / 0.73	0.73 / 0.73	0.73 / 0.73
Model 5	0.77	0.77 / 0.69	0.64 / 0.81	0.70 / 0.74
Model 6	0.7	0.90 / 0.63	0.44 / 0.95	0.59 / 0.76
Model 7	0.74	0.89 / 0.67	0.54 / 0.94	0.67 / 0.78
Model 8	0.74	0.89 / 0.67	0.54 / 0.94	0.67 / 0.78
Model 9	0.98	0.98 / 0.98	0.98 / 0.98	0.98 / 0.98
Model 10	1	1.0 / 1.0	1.0 / 1.0	1.0 / 1.0

Table 5: Evaluation Results

Models/ Metrics	Model Label
BERT fine-tuned with Sentiment140 on Test Set (epoch 5)	Model 1
BERT fine-tuned with Sentiment140 (epoch 5) on Twitter Benchmark	Model 2
BERT fine-tuned with Sentiment140 on Test Set (epoch 10)	Model 3
BERT fine-tuned with Sentiment140 (epoch 10) on Twitter Benchmark	Model 4
BERT fine-tuned with IMDB on Test Set (epoch 5)	Model 5
BERT fine-tuned with IMDB on Twitter Benchmark (epoch 5)	Model 6
BERT fine-tuned with IMDB on Test Set (epoch 10)	Model 7
BERT fine-tuned with IMDB on Twitter Benchmark (epoch 10)	Model 8
BERT Embeddings with RF on Twitter Benchmark	Model 9
BERT Embeddings with SVC on Twitter Benchmark	Model 10

Table 6: Models Evaluated

Due to the computational limitation of the local system and cloud GPUs in this study, we downsized the primary (Twitter Benchmark Dataset) and secondary datasets (Sentiment 140) from over 900,000 records to 14000 records and 1.6 million records to 100,000 records respectively and pre-processed for training and testing on BERT models. We also had to control the parameters for model training phase to avoid system crashes and late runtimes. An exhaustive evaluation of the related models such as DistilBERT, XLNet, NB, KNN etc. and their respective combinations could have provided more concrete grounds for analysis and results. Provided the available computational resources further exploration of the research question can be extended to other NLP tasks such as Emotion analysis and Stance Analysis. Nevertheless, evaluation results from this controlled study proves that the implementation of fine-tuning BERT model approach (approach 1) lags when compared with BERT embeddings used with ML models (approach 2) when it comes to medium to large datasets and incurs high computational resource demand for experimentation, hence it is preferable to use BERT embeddings with an ML model in ensemble to perform sentiment analysis in this problem domain.

5 Conclusion and Future Work

Sentiment analysis is one of the most powerful fields of study that has immense potential for knowledge extraction and utilization in various fields, especially social domain. Businesses and governments can utilize this technology to inspire social change for the better. Social organizations must consider and explore this field of research to gain valuable insights into policy and its impact on society on a regular basis. We implemented a state-of-the-BERT model and fine-tuned it for our tasks and contrasted it with a feature engineered BERT fed into an ML model for classification. We observed that in terms of model complexity, training time, computational resource demand, and model performance, the latter approach (RF-BERT: 98%, SVC-BERT: 100%) outperformed the fine-tuned approach (77%).

Owing to the limited computational resources, the research was done in a controlled environment, however, with the necessary infrastructure larger model design and ensemble techniques can be explored. This research can not only be extended to other domains but also to different social media platforms such as facebook and Instagram for opinion mining. For future work, other state-of-the-art models including roBERTa, DistilBERT, XLNet, XLM etc. will be explored and an exhaustive evaluation will be conducted with the sufficient infrastructure on expansive datasets covering larger demographics to yield results for different age groups and sections of society.

References

Akila, R., BrindhaMerin, J., Vishal, R. and SH, V. K. (2020). Prediction of juvenile delinquencies in correlation with education, *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, pp. 987–993.

Al-Hashedi, A., Al-Fuhaidi, B., Mohsen, A. M., Ali, Y., Gamal Al-Kaf, H. A., Al-Sorori, W. and Maqtary, N. (2022). Ensemble classifiers for arabic sentiment analysis of social network (twitter data) towards covid-19-related conspiracy theories, *Applied Computational Intelligence and Soft Computing* **2022**.

Alammary, A. S. (2022). Bert models for arabic text classification: A systematic review, *Applied Sciences* **12**(11): 5720.

Alatrista-Salas, H., Morzán-Samamé, J. and Nunez-del Prado, M. (2019). Crime alert! crime typification in news based on text mining, *Future of Information and Communication Conference*, Springer, pp. 725–741.

AlBadani, B., Shi, R. and Dong, J. (2022). A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and svm, *Applied System Innovation* **5**(1): 13.

Asghar, M. Z., Khan, A., Ahmad, S., Qasim, M. and Khan, I. A. (2017). Lexicon-enhanced sentiment analysis framework using rule-based classification scheme, *PloS one* **12**(2): e0171649.

Balabantaray, R. C., Mohammad, M. and Sharma, N. (2012). Multi-class twitter emotion classification: A new approach, *International Journal of Applied Information Systems* **4**(1): 48–53.

Bansal, A., Choudhry, A., Sharma, A. and Susan, S. (2022). Adaptation of domain-specific transformer models with text oversampling for sentiment analysis of social media posts on covid-19 vaccines, *arXiv preprint arXiv:2209.10966* .

Bouazizi, M. and Ohtsuki, T. (2019). Multi-class sentiment analysis on twitter: Classification performance and challenges, *Big Data Mining and Analytics* **2**(3): 181–194.

CHIHAB, M., CHINY, M., Mabrouk, N., BOUSSATTA, H., CHIHAB, Y. and HADI, M. Y. (2022). Bilstm and multiple linear regression based sentiment analysis model using polarity and subjectivity of a text.

Chiorrini, A., Diamantini, C., Mircoli, A. and Potena, D. (2021). Emotion and sentiment analysis of tweets using bert., *EDBT/ICDT Workshops*.

Glorot, X., Bordes, A. and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach, *ICML*.

Hassan, T., Bajwa, I. S. and Hassan, S. (2016). Prediction of terrorist activities by using unsupervised learning techniques, *Journal of Applied and Emerging Sciences* **6**(2): pp56–60.

Hossain, M. S. and Rahman, M. F. (2022). Customer sentiment analysis and prediction of insurance products' reviews using machine learning approaches, *FIIB Business Review* p. 23197145221115793.

Kazmaier, J. and van Vuuren, J. H. (2022). The power of ensemble learning in sentiment analysis, *Expert Systems with Applications* **187**: 115819.

Kodiyala, V. S. and Mercer, R. E. (2021). Emotion recognition and sentiment classification using bert with data augmentation and emotion lexicon enrichment, *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 191–198.

Kumar, J. S., Sri, J. D., Manoj, K. M. S., Srividya, D., Prathyusha, D. and Kumar, M. S. (2022). Sentiment analysis on textual tweets using ensemble classifier (lstm-gru), *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, IEEE, pp. 926–932.

Marcec, R. and Likic, R. (2022). Using twitter for sentiment analysis towards astrazeneca/oxford, pfizer/biontech and moderna covid-19 vaccines, *Postgraduate Medical Journal* **98**(1161): 544–550.

Mohammad, S. M., Kiritchenko, S. and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets, *arXiv preprint arXiv:1308.6242* .

Mohd, M., Javeed, S., Nowsheena, Wani, M. A. and Khanday, H. A. (2022). Sentiment analysis using lexico-semantic features, *Journal of Information Science* p. 01655515221124016.

Monika, R., Deivalakshmi, S. and Janet, B. (2019). Sentiment analysis of us airlines tweets using lstm/rnn, *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, IEEE, pp. 92–95.

Mukhtar, N., Khan, M. A. and Chiragh, N. (2018). Lexicon-based approach outperforms supervised machine learning approach for urdu sentiment analysis in multiple domains, *Telematics and Informatics* **35**(8): 2173–2183.

Naseem, U., Khan, S. K., Razzak, I. and Hameed, I. A. (2019). Hybrid words representation for airlines sentiment analysis, *Australasian Joint Conference on Artificial Intelligence*, Springer, pp. 381–392.

Naseem, U. and Musial, K. (2019). Dice: Deep intelligent contextual embedding for twitter sentiment analysis, *2019 International conference on document analysis and recognition (ICDAR)*, IEEE, pp. 953–958.

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W. and Kim, J. (2021). Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis, *IEEE Transactions on Computational Social Systems* **8**(4): 1003–1015.

Ombabi, A. H., Ouarda, W. and Alimi, A. M. (2020). Deep learning cnn–lstm framework for arabic sentiment analysis using textual information shared in social networks, *Social Network Analysis and Mining* **10**(1): 1–13.

Phu, V. N., Tran, V. T. N., Chau, V. T. N., Dat, N. D. and Duy, K. L. D. (2017). A decision tree using id3 algorithm for english semantic analysis, *International Journal of Speech Technology* **20**(3): 593–613.

Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M. and Baz, M. (2022). Transfer learning for sentiment analysis using bert based supervised fine-tuning, *Sensors* **22**(11): 4157.

Sahar, A., Ayoub, M., Hussain, S., Yu, Y. and Khan, A. (2022). Transfer learning-based framework for sentiment classification of cosmetics products reviews, *Pakistan Journal of Engineering and Technology* **5**(3): 38–43.

Shamrat, F., Chakraborty, S., Imran, M., Muna, J. N., Billah, M. M., Das, P., Rahman, O. et al. (2021). Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm, *Indonesian Journal of Electrical Engineering and Computer Science* **23**(1): 463–470.

Singh, J. and Tripathi, P. (2021). Sentiment analysis of twitter data by making use of svm, random forest and decision tree algorithm, *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, IEEE, pp. 193–198.

Sinha, S., Jayan, A. and Kumar, R. (2022). An analysis and comparison of deep-learning techniques and hybrid model for sentiment analysis for movie review, *2022 3rd International Conference for Emerging Technology (INCET)*, IEEE, pp. 1–5.

Subba, B. and Kumari, S. (2022). A heterogeneous stacking ensemble based sentiment analysis framework using multiple word embeddings, *Computational Intelligence* **38**(2): 530–559.

Suresh, H. et al. (2016). An unsupervised fuzzy clustering method for twitter sentiment analysis, *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, IEEE, pp. 80–85.

Tang, D., Qin, B. and Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification, *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1422–1432.

Tang, D., Wei, F., Qin, B., Liu, T. and Zhou, M. (2014). Coooolll: A deep learning system for twitter sentiment classification, *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pp. 208–212.

Umer, M., Ashraf, I., Mehmood, A., Kumari, S., Ullah, S. and Sang Choi, G. (2021). Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model, *Computational Intelligence* **37**(1): 409–434.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.