Research Project

Gender Prediction Based on various Nationality Names using Deep Learning techniques.

Deepak Kumar

Research Project
Supervisor : Prashanth Nayak
National College of Ireland
IFSC, Dublin

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Deepak Kumar……

Student ID: X20195028……

Programme: M.Sc Data Analytics…… **Year:** 2022……

Module: Research Project……

Supervisor: Mr. Prashanth Nayak……

Submission Due Date: 01/Feb/2023……

Project Title: Gender Prediction Based on various Nationality Names using Deep Learning techniques

Word Count: …6326…… **Page Count** …22……

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: Deepak Kumar……

Date: 29/Jan/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):	

Abstract

Gender Prediction task has been carried out in this paper for various nationality names. Gender prediction by name was not a simple task to implement. Gender is an attribute of demographic data of individuals which plays a significant role in identifying patterns and analysing the industrial, business, and medical data to plan the strategies for the market. Previously, many authors have carried out work in the same domain to predict gender based on the name of different languages and different nationalities. However, this paper also deals with gender prediction of different nationalities using text classification which is different from others as no one has used the sequence classification technique to predict gender. In this paper, the name of individuals has been considered and classified using different sequence classification technique and trained by multiple deep learning techniques. An open-source dataset acquired from UCI is used for text classification. DistilBERT, BERT, RoBERTa and XLNet are the four model which has to implement in this paper for analysis of prediction and comparison proved DistilBERT model as the most efficient model based on results obtained by computing metrics and time.
Keywords : Gender Prediction, Deep Learning, DistilBERT, Multi-Class Text Classification

Contents

List of Figures

List of Tables

Introduction

In today's era, technology has stepped into almost every domain. With the fast pace of technology, artificial intelligence has evolved to simplify the problem for mankind. In every field of technology, artificial intelligence proved to be the best way of estimation and prediction. However, generating the inference is not that much simple as it sounds in most of the scenarios. To generate precise and accurate prediction machine learning and deep learning techniques can be applied.

1.1 Background of Deep learning for Gender Prediction

There is some web application available in the market which can predict a gender by name but they are limited to several names which are already set in their dictionary (To et al. 2020, p. 55). In case any name out of their dictionary is requested such applications are failed. This limitation of the domain attracted the attention of researchers and then authors started finding a solution for it. Also, this limitation motivated the work carried out in this paper. So this paper tends and aimed to identify the naming features and patterns responsible for male and female names. This work would be able to contribute to multiple domains for example gender of patients is missing from the database of a hospital or due to incorrect data entry and gender is considered as significant data as the diagnosis or any clinical trial is completely based on it. In many surveys or forums, it is not suggested to ask about the gender of the user and some users don't feel comfortable revealing their gender. Many outlets and stores have the data shopping patterns items but at the billing, they do not enter the gender. In such cases, the analysis of data can be challenging and may not be that much accurate to overcome such issues approach of this paper can help to overcome the problem of the domain.

1.2 Objective of Research

The primary objective of the research is to predict gender as it will be helpful in multiple domains. The contribution of the study is to find gender using sequence classification. By using N-gram[1] and similar packages it was required to consider more than three alphabets

[1] `https://pypi.org/project/ngram/`

as those can lead to over-fitting but the sequence classification tends not to miss the naming feature with less than three alphabets. This would add an advantage and lead to less miss-classification and to generate a more efficient model (Ho Huong et al. 2022, p. 3). Many authors paid attention to this domain but none of them has implemented the sequence classification technique. Also from previous work, it is concluded that the first name is most significant for the gender as the last name is least significant for the identification of gender. This work can also help in the preprocessing part of the studies where the gender-related data is missing or unavailable. (Rego & Silva 2021, p. 1) also mentioned the complete data is mandatory for the inference models, training and techniques to implement the analysis using artificial intelligence. The technique we have implemented Sequence classification will identify all the naming features in the name even if any pattern is found for three or less than 3 alphabets which was ignored by most of the authors in the previous work.

RQ: How well can Sequence Classification applied to the various nationalities names of individuals to predict the gender using deep learning Techniques?

Sub RQ: Can Sequence classification techniques with multiple nationality names classify the gender without over-fitting?

To develop the understanding of the research question and sub research question below mentioned objectives have been carried out, achieved and interpreted.

1. Identify the scope of work, exploration and critical review of the previous work carried out in the last five years for gender identification by name.

2. To develop and design the framework for the approach that can clearly explain the implementation and results of the research for Gender Identification.

3. Implementation of the exploratory data analysis and pre-processing of the data-set before model training.

4. Feeding the dataset to model for the implementation, and evaluation followed by a discussion of the results for all four models used in this paper.

    4.1. Execution, compute metrics and Interpretation of DistilBert.

    4.2. Execution, compute metrics and Interpretation of Bert.

    4.3. Execution, compute metrics and Interpretation of Roberta.

    4.4. Execution, compute metrics and Interpretation of XL Net.

5. Detailed comparison of the implemented four models and ruling out the best model.

In this paper, multiple text classification deep learning models have been implemented using the sequence classification technique which has been referred from the hugging face. the structure of the report will be followed as per the reporting standards. The introduction section has completely covered the problem domain, research question and the objective of

the research. It has been followed by a literature review where a detailed explanation supports the choices of techniques and packages used for the implementation. The methodology section has elaborated on the steps of pre-processing, steps of models training and results. In last the results are discussed in detail to understand and justify the best model.

Related Work

Multiple authors (Ani et al. 2021, p. 2), (Ali et al. 2016, p. 161), (Tripathi & Faruqui 2011, p. 137), (Jia & Zhao 2019, p. 676) focused on single language-based name gender identification for Hindi, English, Arabic, Bangladesh, Japanese, Vietnamese, etc. Especially in the Asian names, a common pattern was considered similar for most of the names. This pattern was the morphological analysis of the names considered. Like most female names are ending with vowels, based on the number of syllables[1] present in the name, Sonorant consonant[2], and length of the name.

From the view of morphology,(Jia & Zhao 2019, p. 676) aimed to identify the gender of the Chinese names. However, the formation of Chinese names is different from south Asian names. But the author said the east Asian names as logosyllabic[3]. According to (Sun & Huo 2022, p. 2) Using the Pinyin approach Chinese names were embedded and trained using the BERT model. As BERT supports multiple languages, it showed the best performance with 93 per cent than other techniques like NB, and Gradient Booster which were trained using FastText.

Apart from the name, gender identification of an individual by iris recognition, face image classification has already been carried out. But the most interesting work carried out by the (Shrestha et al. 2016, p. 3394) was identifying the gender by the posts of the member in a social medical group named "Daily Strength". The author scratched data from the member's post, reposted and replied to the post. As it was medical support group 65 per cent of the post, the reply had gender-revealing information since author profiling was carried out for these posts. Ngram package was used for the classification of the familial token where up to 3 words were considered. Posts like my wife, my husband, and my daughter were revealing the gender itself. All these tokens were transformed to feature and trained to machine learning and deep learning models. Out of multiple techniques, Logistic regression shows the best results with 88.29 per cent of accuracy.

[1] https://dictionary.cambridge.org/dictionary/english/syllable

[2] https://allthingslinguistic.com/post/68721010548/detailed-explanation-sonorants-obstruents-sonority

[3] https://glosbe.com/en/en/logosyllabic

2.1 Distilbert and BERT for Text Classification

The DistilBert model [4] is considered as fast, small, and efficient and belongs to the transformer model which is trained by the fine BERT model. It is considered to be more efficient than the BERT model and more simple to use and almost 65 per cent faster than BERT as benchmarked based on GLUE[5] language.

The (Dolci 2022, p. 3) has used primarily BERT and Distilbert models for natural language processing where they showed the probability of mitigating the gender using finetuning the sentence based on similarity built among the gender-centred sentences and gender-swapped sentences. However, using a small dataset author concluded the satisfactory results that mitigation techniques are helpful to decrease the gender bias for the sentence encoders.

In another paper, (Qasim et al. 2022, p. 6) claimed that natural language processing and data mining has attracted the attention of researchers worldwide to develop an automatic system for text classification. The author has identified and classified the fake news from the tweets related to COVID-19 in another dataset and classified the extremist and non-extremist tweets by tokenizing the words like suicide, ISIS, bomb, etc. In the last dataset, the author identified informative tweets versus uninformative tweets for Covid-19. Uninformative tweets were categorised and refined with less than 9 words, removal of tweets was for the person who has less than 5 followers, and tweets were not retweeted. For all three datasets scratched and multiple text classification techniques were implemented including Bert-base, Roberta, Distilbert, Albert, etc.

A comparison of all 9 models showed and later discussed concluded that BERT achieved 98.83 per cent of accuracy for English tweets datasets, BERT outperformed with 99.71 accuracy for extremist and non-extremist datasets and last, Roberta-base resulted in 99.71 per cent of accuracy for the fake news related to Covid-19. Since results and art of this paper motivated to carry out the work on name gender identification as these text classification models are much better than machine learning and other deep learning techniques as previously discussed the Fast text and N-gram packages used to tokenize the alphabets and then training the models like NB, LSTM, SVM, etc.

In another research, (Soldevilla & Flores 2021, p. 2) stated that the posts from social media can be classified to identify the violence against women. Reddit and Twitter were considered the primary platforms to scratch the posts related to violence against women. The author performed fine-tuning processing to the BERT model to classify the label. this dataset was split into train, validation and test set and later evaluated and compared with different epochs. In this study, the BERT model was not compared with any other model instead result of it were compared with different training arguments. The outcomes of the study for BERT were 87.41 per cent of accuracy with 6 epochs and negligible difference with 87.30 per cent of accuracy with 3 epochs. It was concluded that fine-tuned BERT can be applied to other text classification problems as it gave an AUC of 0.9603 and was able to classify the violent and non violent posts against women.

[4]Distilbert referred from `https://huggingface.co/docs/transformers/model_doc/distilbert`

[5]`https://gluebenchmark.com/`

2.2 RoBERTa and XLNET for Text Classification

A variety of work has been carried out for text classification of Chinese language. As mentioned above that BERT has already been implemented for the identification of gender for Chinese names using different machine learning and deep learning techniques which were trained by FastText in 2019 by (Jia & Zhao 2019, p. 676). After two years in the same domain of the text classification of the Chinese language the FastText package by replaced by fine-tuned model RoBERTA which is a more refined version of the BERT. In this paper, (Xu 2021, p. 3) has mentioned and compared the work of BERT and RoBERTa and added it to the progress of neural networks in the field of Natural Language processing in past years. A pilot dataset for the study was considered from Susong[6]. Data were manually labelled 1 for illegal behaviours and rest of them with 0. Around 6755 samples were processed and split into training and testing for the training of the models. RoBERTa was fine-tuned with multiple models like CNN (Convolutional Neural Network), RNN (Recureent Neural Network), DPCNN (Deep Pyramid Convolutional Neural Networks) and RCNN (Region-Based Convolutional Neural Network) with 10 epochs. A detailed comparison of all these models was carried out with the classification of baseline RoBERTa and concluded that none of the models except RNN was able to come even close to a baseline model. (Xu 2021, p. 3) showed that the RoBERTa-wwm-ext were the best among the others and called the strongest fine-tuned model.

After considering the multiple article and research papers, the approaches used in the above-mentioned paper work to be carried out is motivated to implement text classification using fine-tuned models. Transformers is a widely used and common deep-learning library for python. And this paper has acquired implementation using transformers and multiple fine-tuned models kept in consideration for this paper. Another work for Roberta using transformers for text classification was carried out by (Alam et al. 2020, p. 4) for the Bangla language which is an Asian language. One more thing to consider is that more than 100 languages are compatible with this package which also made it common and made it robust. Multiple languages on a platform are classified which made research easy to implement, no different packages for different languages are required. Authors have also mentioned in the past few years for semantics and word embedding different neural networks like Multi-perceptron, Recurrent Neural Networks or Convolution Neural networks were used but in the most recent research, transformers-based pre-trained models are successfully used for learning the language representation for a large amount of data. The authors considered multiple datasets from youtube comments, News comments, etc which were used previously with deep learning techniques and performed the experiment with two models i.e. BERT and XLNet. And concluded that XLNet have performed better among all classical machine learning techniques. Even as compared with BERT and evaluated relatively better with 8 per cent for Youtube sentiment Analysis, 21 per cent better for YouTube emotion analysis and a minor improvement with 1 per cent for News classification.

Overall, it can be easily concluded that the for the domain of text classification the finetune and pre-trained models can classify gender efficiently as compared to other machine learning models. Since none of the machine learning models is implemented in this paper.

[6] https://susong.tianyancha.com/

Deep Learning models like LSTM and ANN has shown better results but they are based on text classification package for embeddings and tokenizing the features to train the model.

2.3 Finetuned models over Text classification packages

A variety of text classification packages are available that help in tokenizing the text so that any machine learning or deep learning model can be trained. Most of the authors in the past 5 years have recommended the machine learning techniques like Support Vector Machines (SVM) and Naive Bayes (NB). For Deep learning most common were Long Short Term Memory (LSTM), CNN and ANN. But in the past years either the work carried out for name or any research related to text classification has primarily used the fine-tuned model using transformers. This also motivated the work in this paper using a fine-tuned model.

2.3.1 N-gram Packages

Many text classification packages were evaluated and many classifiers were kept in consideration before implementation of Machine and Deep learning techniques as it was the way to classify the text. In the past 2 years fined tuned models are evolved and almost replaced the traditional way of classification. Ngram package has been used and stated by (Ho Huong et al. 2022, p. 3) and (Shrestha et al. 2016, p. 3394) as efficient and faster but the limitation is also to classify the familial token considers up to uni, bi or tri grams else it will lead to overfitting which was the primary drawback of using Ngrams. An overview for the Ngrams models is displayed in Fig.2.1 On the contrary models from transformers is capable of auto-tokenising and auto-sequencing the text which makes it feasible, short, easy to use and faster. Every model BERT, DistilBERT, ROBERTa, XLNet are capable and can perform the text classification task independently.

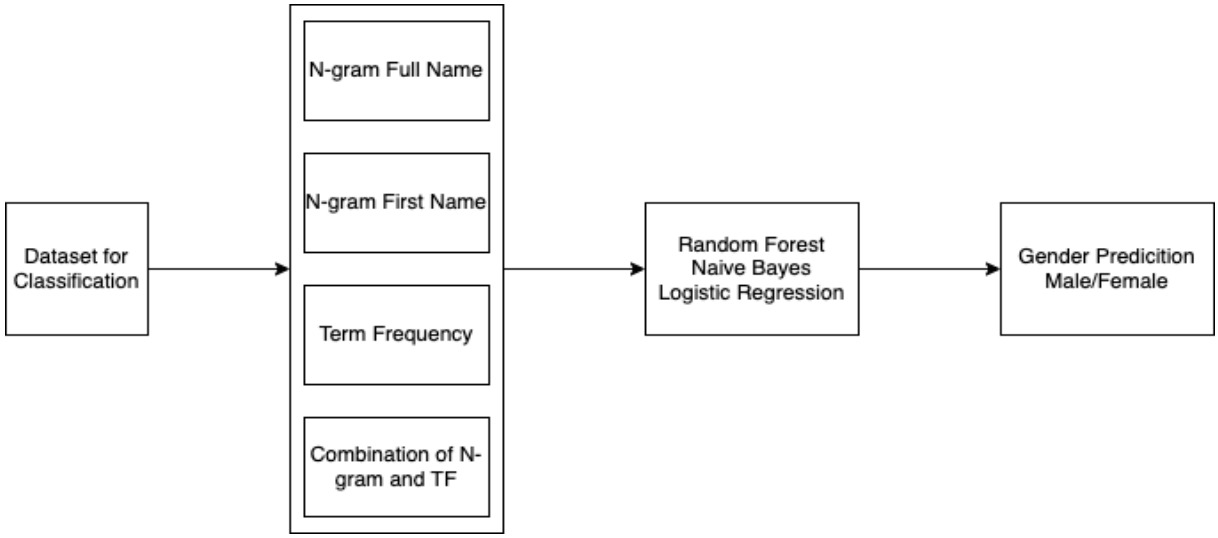

Figure 2.1: Overview of architecture using N-Gram package
(Ho Huong et al. 2022, p. 3)

2.3.2 FastText

From (To et al. 2020, p. 55), (Jia & Zhao 2019, p. 676), (Mahendra et al. 2022, p. 131) it is concluded that FastText works well with deep learning techniques. In the same way for the approaches using the FastText class many authors carried out work with this and FastText has efficiently proved its performance on Vietnamese names dataset known as UIT-ViNames. Where LSTM and FastText outperformed with more than 95 per cent of accuracy. In another paper on the Chinese language author classified text using FastText. FastText has the advantage over Ngram in that it supports more than 70 languages worldwide and is capable of classification for most languages. However (Grave et al. 2018, p. 1), said it works around 157 languages with a batch size of 300 and 2 epochs. This paper mentioned classification using FastText can boost the efficiency of the model, especially for Deep learning techniques. In the same way, many other text classification package is available like Word2vec, Bag of words, Genderize, etc. and authors has implemented them and gained somewhat better results. But for the many other text classification problems, these models have failed.

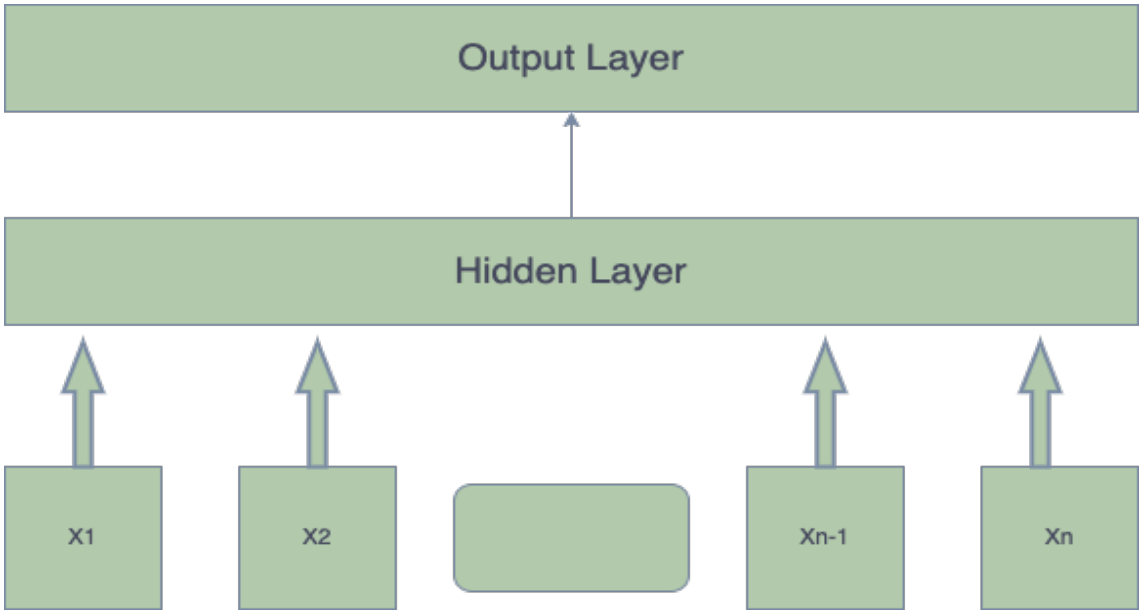

Figure 2.2: FastText working Model
(Yao et al. 2020, p. 5)

2.3.3 Comparison Table of ML and Deep Learning

It can be seen that most of the studies were based on traditional machine learning techniques or deep learning with text classification packages. Additionally, in recent studies, many researchers have encouraged and motivated the use of fine-tuned models for text classification. From past work, implementation using sequence classification for the prediction of gender by name is unexplored which shows the novelty and improvement.

There is still enough to explore and a lot of previous work. It's not possible to refer to all of them but the most significant and useful study is kept in consideration and referred to above. But some results from previous studies are shown below in the table which gives a

clear understanding of successful models and those models which were failed. From a paper (Çoban et al. 2021, p. 3), the useful table for text classification models with results has been acquired and displayed below where information on Datasets, Text classification package, methods and results are displayed.

TF package	**Users Set**	**Language**	**Method and Results**
Bag of Words	5200 Tweets	Turkish	SVM with 72.3 percent
N-grams and bag of words	3 datasets of PAN@CLEF	English, Arabic, and Spanish	SVM with 82.7 percent
Bag of Words	profiles scratched from 3000 posts from users	Language independent	NB and RF with 63.0 percent
Words referring gender	PAN'16 tweets collected	English	Basic Linear Classifier with 61.0 percent
Word embeddings	2400 users tweets	Arabic	Primarily LSTM and CNN with 79.6 percent
Bag of words	8700 replies of tweets out of 3000 tweets	Turkish	LR with 74.1 percent
Word embeddings	Tweets of 4000users	Turkish	ANN with 80.6 percent
Features with activities and images	1.2M tags and image data of 3700 users	English	Primarily SVM and MLP with 82.0 percent

Table 2.1: Table for previous researchers for ML and Deep learning techniques (Çoban et al. 2021, p. 3)

Methodology

This section of the paper provides detailed information of its design, problems that arise during the research and how were they addressed. This section address objective number 3 & 4. Below mentioned design of the study is based on the implementation and discussed in upcoming sections.

3.1 Data Mining

For the classification of gender based on names, CRISP-DM (Cross Industry Standard Process for Data Mining) technique has been optimised and then executed and implemented for it. In the figure 3.1, the overview of the methodology visualised the phases of the project. Additionally, it gives the right way, the right procedures and a foundation for improved and quick results. CRISP-DM method is used widely and has acceptable standards for a project, especially in any Data related projects which also accounts for the benefits of CRISP-DM. The life-cycle of this model for this project has six phases as shown in figure 3.1 which are Problem Domain and objectives, EDA (Exploratory Data Analysis), Pre-processing and transformation, Proposed Models, Model Evaluation and Results. The actual steps of a CRISP-DM may differ from this optimised model.

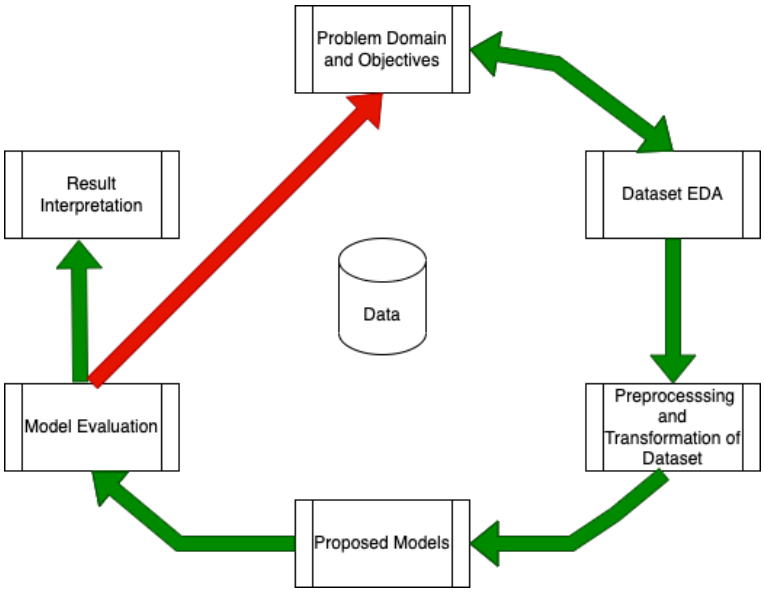

Figure 3.1: Overview of Methodology used for gender prediction by name
What is CRISP DM? - Data Science Process Alliance (n.d.)

3.2 Planning Phases

Whenever it is about a project or business, primarily it is required to understand the objectives. Mainly there are four tasks for a project or goals, firstly to address the objectives where and how to focus the real objectives of the project i.e classification of gender based on the name for this paper. Secondly, specify if any assumptions are made or required before or during the project. Next is to identify the objective of classification which refers to labels in this paper. Lastly, to give a good project plan, techniques, dependencies and tools required to achieve the primary objectives.

3.3 Data Description and EDA

The dataset used for this research project has been acquired from open source platform (Dua & Graff 2017, p. 0) UCI machine Learning repository[1]. This dataset is open source and licensed to be used for educational purposes. The licence of the dataset has been provided in the ethics form submitted. This dataset has four general attributes Count, probability, and Name including Gender. Count and Probability have been disregarded as it has zero significance over predicting a gender by name.

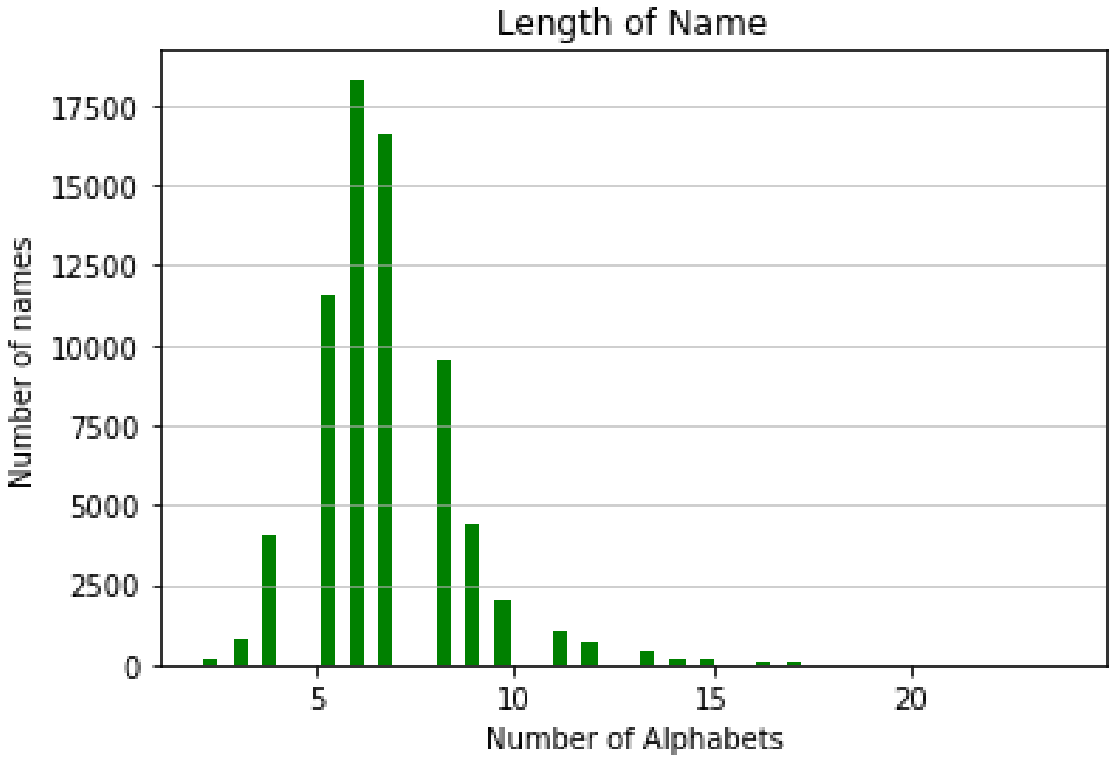

Figure 3.2: Length of the Names

Out of 70000 names, more than 43000 are female and around 27000 are male for the train set for the test set out of 1000 names 590 are female and 410 are male as shown in the figure. These names and gender are originally acquired from four countries' birth registration centres of UK, Australia, America and Cannada. Including the minor class imbalance, the length of the name is also visualised to understand how to sequence classification can work for more than 3 characters and does not lead to overfitting.

[1] https://archive.ics.uci.edu/ml/datasets/Gender+by+Name

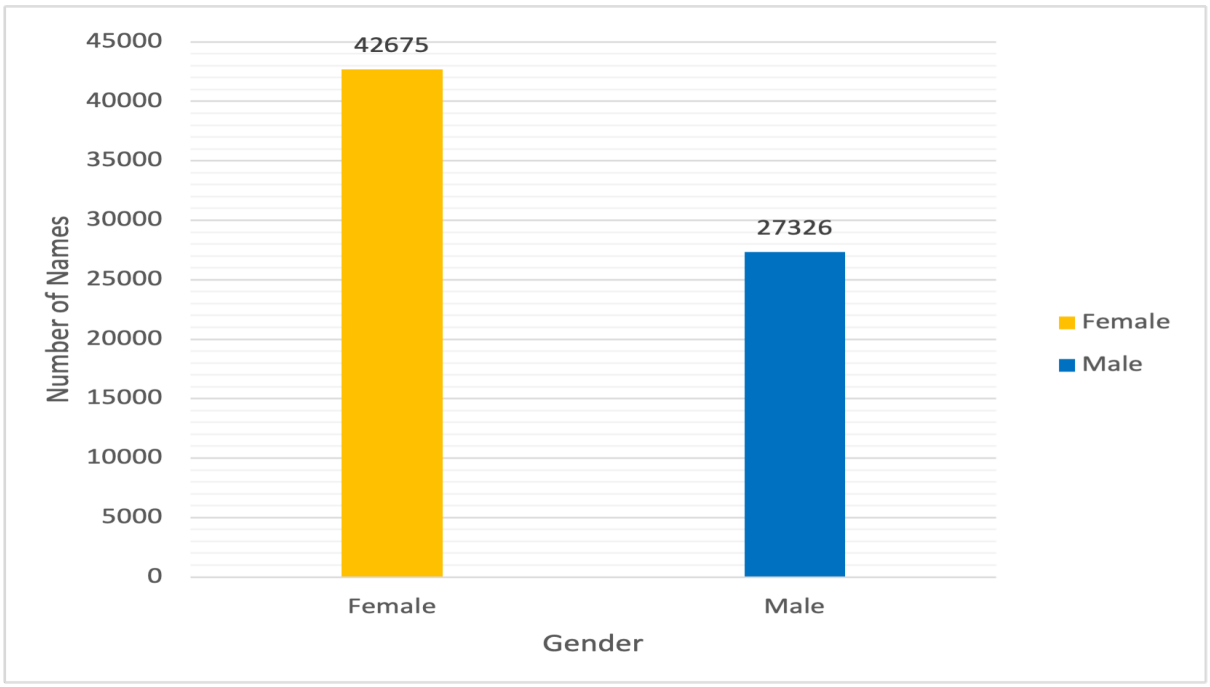

Figure 3.3: Ratio of Male and Female in Train set

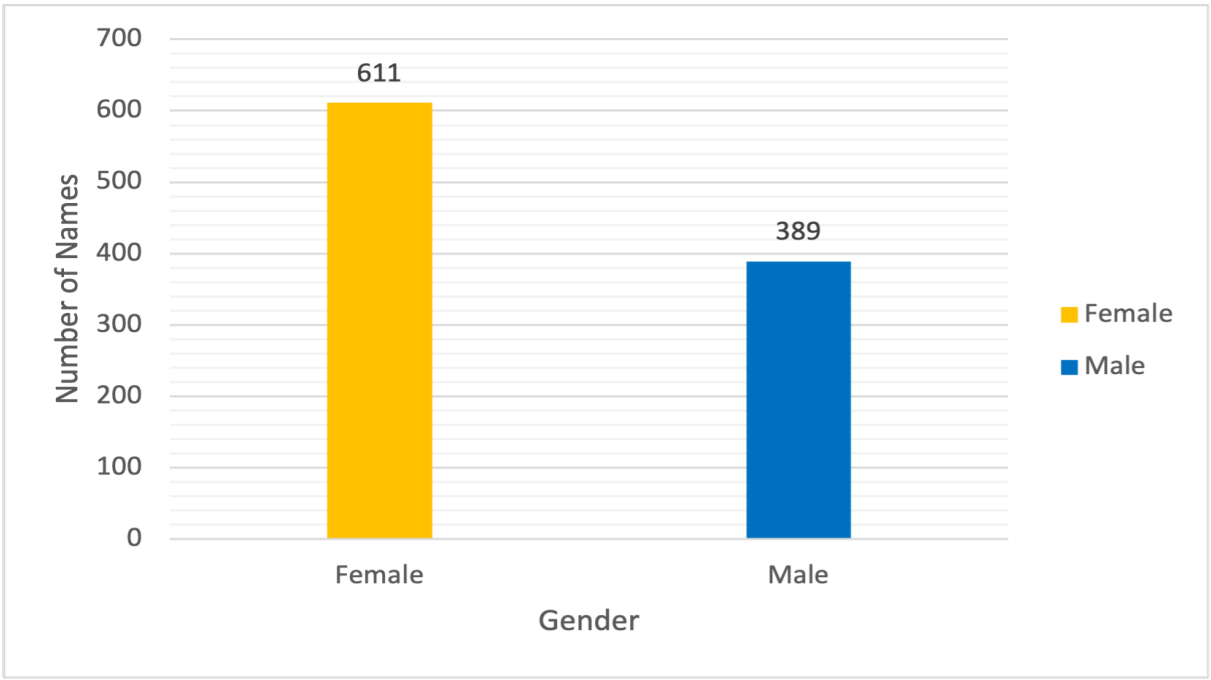

Figure 3.4: Ratio of Male and Female in Test set

3.4 Data Pre-processing

For this section, after ensuring that this dataset suits and shuffling the sequence of names as they were provided in alphabetical order it is considered to be ready for modelling. For the transformation of the dataset, multiple operations have been applied to it. Even preprocessing of the dataset is considered to be more time-consuming than model building. However, for this dataset preprocessing was finished comparatively earlier, but model building and model training took more time. Some operation of data transformation was checking the entire Name column for duplicate values and removing the existing duplicate items. Before removing such duplicate names also checked the gender difference because for the same name different gender could have been considered as a unisex name, but there were zero entries for it. After that checked both columns' Name and Gender for null values if any one of them is missing entire row has been removed. Ensured that all the characters are transformed to lowercase. Additionally, alphabets from a-z have been considered for the name any special characters like (,./';) have been disregarded from the name. The gender column has been encoded in integer value where 1 denotes female and 0 denotes male. However, Lemmatization has not been considered as the different spelling may change the gender else for any text classification lemmatization is recommended.

3.5 Model Selection

In this section of this paper, it is elaborated that four models have been considered for the model building which is BERT, DistilBERT, RoBERTa and XLNet. All of these models are fine-tuned models and have been referred from the hugging face open source platform for machine and deep learning techniques. All these techniques were inspired after critically reviewing the previous research papers for the approaches and their outcome. Based on this as mentioned in the literature review section fine-tune models has been evolved in the past 2 years. Another factor for the selection of all these techniques was that they support the model-based sequence classification method. After fitting the dataset with the model using compute metrics function the outcome of all models has been compared, evaluated, interpreted and then visualised. A detailed discussion of the model is provided in the next Section.

Experiments

This section of the paper discusses in detail the implementation, interpretation of results and visualisation of the proposed four models which are BERT, DistilBERT, RoBERTa and XL-Net. Evaluation of the model is not only carried out considering the training and test results

but also based on similar predictions done and compared with actual values as provided in the test set. After Considering the knowledge from the literature review to proceed with the models. And before that, the visualisation made it easier to understand the data to check what differences are there in name and gender. In the end of this section, a comparison of all four models has discussed based on the results shown in this section.

4.1 BERT (Bidirectional Encoder Representations from Transformers)

First of all, the Deep learning model that has been used for the classification is BERT specifically this model BERT base model (uncased), This version [1] of the model is primarily used against others because for most of the preprocessing steps like converting the alphabet to lower case or removing punctuation are carried out for this model by default. BERT is a pre-trained model of transformers recommended for giant textual datasets and does not require any human intervention to label data generated from input texts. With significant changes, more than 24 models are released after it. It is a bi-directional trained model that does not need to be left to right or right to left training. Classification of text is done very similarly to NSC(Next Sentence Classification)[2], which is integrated by a classification layer on the top of the transformer Output for the token. The larger the data for training means more training steps which help to achieve higher accuracy.

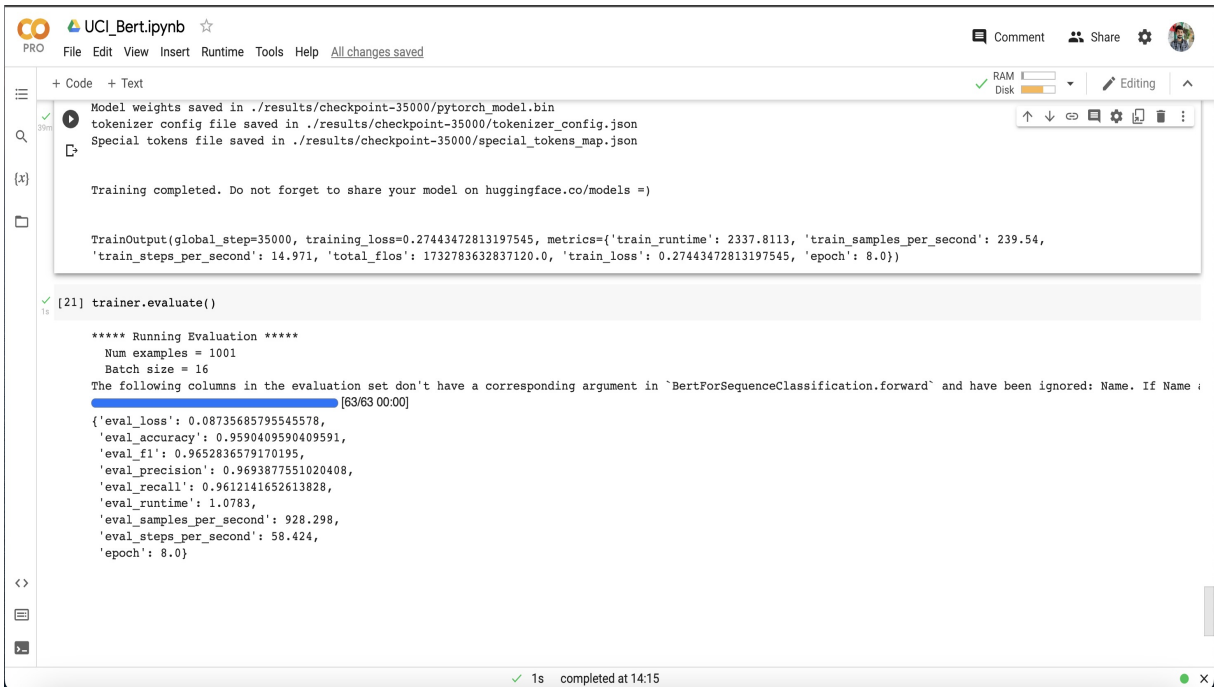

Figure 4.1: Evaluation Results for the BERT base uncased model

[1] https://huggingface.co/bert-base-uncased

[2] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

A variety of training arguments are available which are based on input and mechanism required here learning rate has been kept as 2e-5(0.00002), and the batch size for the train and test set is 16. As for comparing and identifying the baseline model all models are trained with 8 epochs with 0.01 weight decay and evaluated the performance of the model using compute metrics function. From compute metrics function, requested the values of Accuracy, F1 Score, Precision and Recall. In the above figure 4.1 it is shown the model performed quite well with 95.90 per cent of accuracy, the F1 score evaluated is 96.52 per cent, Precision and Recall are almost equal with 96.93 per cent and 96.12 per cent respectively which also show there is no class imbalance factor affect.

4.2 DistilBERT (Distilled Version of BERT)

Secondly, another deep learning model that has been implemented for the text classification is DistilBERT base model (uncased)[3] , this model is a distilled version of the BERT base model. Additionally, this version of the model has an advantage over others as some of the preprocessing steps like converting the alphabet to lowercase or removing punctuation are carried out for this model by default. It is simply a common function of the uncased model. DistilBERT is a pre-trained model of transformers suggested for the self-evaluation as BERT base is its base model and considered as BERT acts like a teacher for it. This model as well does not require any human intervention to label data generated from input texts. There are now revised or improved versions for it has been released yet. In this paper, the Pytorch python package has been used for implementing and importing the dependent package. This model is considered a smaller, cheaper, faster and quick Transformers model. It has an advantage[4] over BERT's speed as it is faster and capable to handle big data like million of million rows which are considered almost 24 times bigger data and can be trained and processed faster. DistilBERT is an efficient model and well compared with BERT for performance as well as almost 95 per cent performance is the same with 40 per cent fewer parameters.

[3] https://huggingface.co/distilbert-base-uncased

[4] https://medium.com/huggingface/distilbert-8cf3380435b5

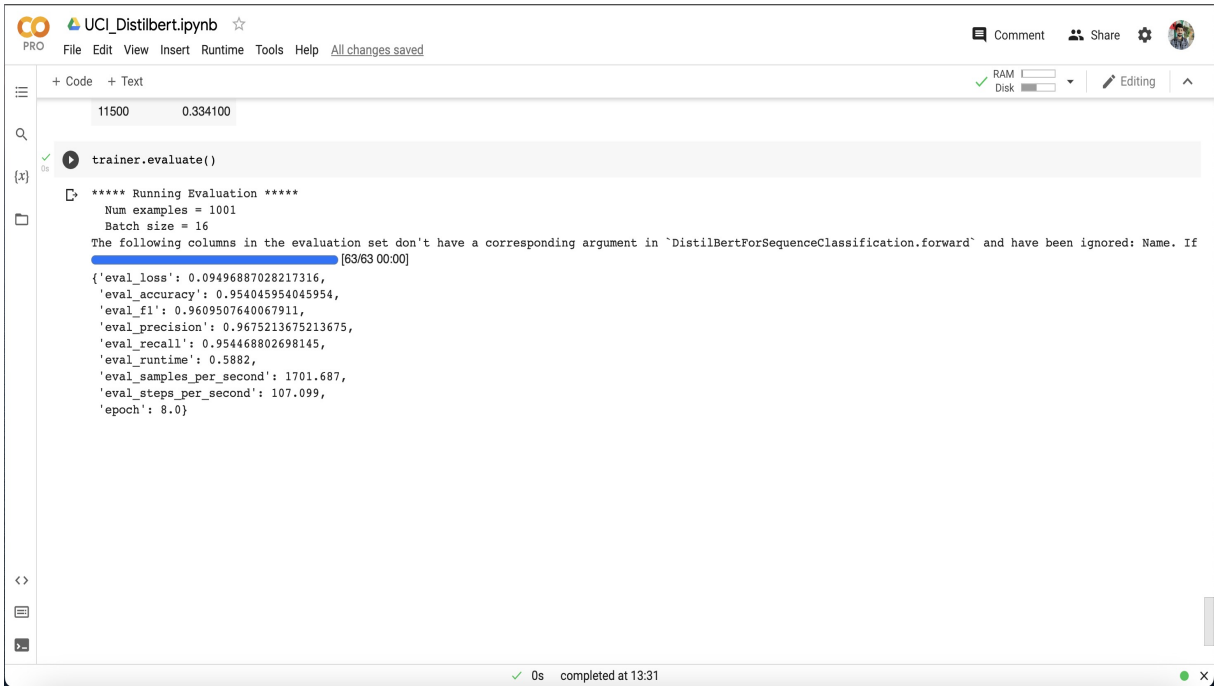

Figure 4.2: Evaluation Results for the DistilBERT model

Training arguments are completely based on the input and mechanism required and the parameter are kept the same for final comparison. Since the learning rate has been kept as 2e-5(0.00002), and the batch size for the train and test set is 16. As for comparing and identifying the baseline model all models are trained with 8 epochs with 0.01 weight decay and evaluated the performance of the model using compute metrics function. From compute metrics function, requested the values of Accuracy, F1 Score, Precision and Recall. In the above figure 4.2, it is shown the model performed quite well with 95.40 per cent of accuracy, the F1 score evaluated is 96.09 per cent, and Precision and Recall are almost equal with 96.75 per cent and 95.44 per cent respectively which also show there is no class imbalance factor affect here as well.

4.3 XLNET

After DistilBERT, the third model for the evaluation for this work is XL Net. It is a Generalised autoregressive pre-trained model used for language understanding[5]. It is an unsupervised text representation learning method and it is based on the combination of text modelling objectives. In addition, It comes with the backbone of transformers, which is proven as an excellent performing model for language/text-involved tasks. This model is primarily fined tuned used for the downstream task. It aims to use the entire text to predict using sequence classification and token classification. However, this model has mostly been used to predict further phrases and sentences. It is tried to classify the gender based on name as it has huge capability in text classification and prediction for future required

[5]https://huggingface.co/xlnet-base-cased

words. This model is not uncased hence it is case-sensitive and does not ignore any special characters/ punctuation. However, pre-processing steps are explained above and only after those operations this model is implemented.

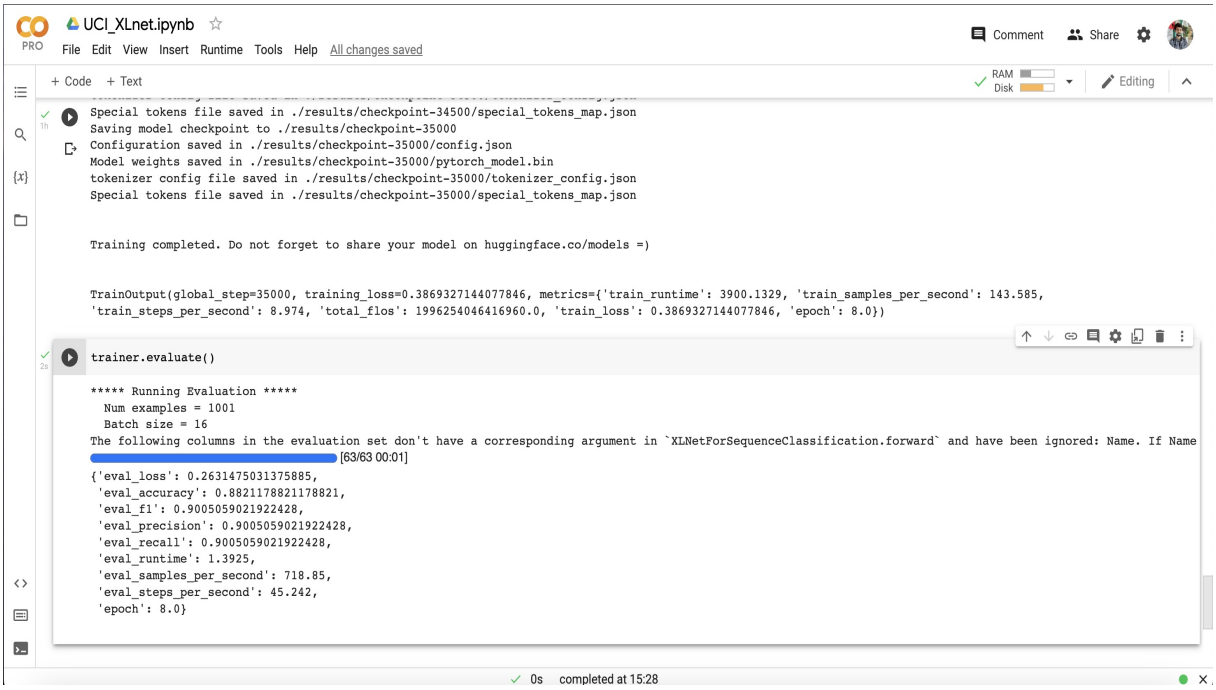

Figure 4.3: Evaluation Results for the XL Net model

Again, the training arguments have been kept the same for final comparison. Including learning rate and all other parameters are the same. The learning rate is 2e-5(0.00002), and the batch size for the train and test set is 16. As for comparing and identifying the baseline model all models are trained with 8 epochs with 0.01 weight decay and evaluated the performance of the model using compute metrics function. From compute metrics function, requested the values of Accuracy, F1 Score, Precision and Recall. In the above figure 4.3 it is shown the model performed quite well with 88.21 per cent of accuracy, the F1 score evaluated is 90.05 per cent, and Precision and Recall are equal with 90.05 per cent and 90.05 per cent respectively which also show zero signs of class imbalance.

4.4 RoBERTa (Robustly Optimized BERT Pretraining Approach)

Lastly, implemented the fourth model is RoBERTa[6], it also belongs to transformers' pre-trained model and is trained on a large dataset for the English language in a self-supervised manner. In the same way, as DistilBERT does not require any human intervention it also labels the input text itself but it's different because it supports the English language. It is intended to use as a fine-tuned model for a downstream task. It is primarily objected to

[6]https://huggingface.co/roberta-base

fine-tuning the task as phrase/sentence to decide the embedding/tokenization like sequence classification or token classification. The approach of this model[7] is to learn the inner representation of text and then extract features from it to downstream the task. The backend working of this model is similar to BERT. Training arguments as seen below are also same for this model implementation. The results of this model can be seen from the below figure 4.4 as 91.8 per cent of accuracy with a 93.07 per cent of F1 score. Whereas the Precision value is 93.23 per cent with quiet a closer value of recall i.e. 92.91 per cent. These results for RoBERTa are above the range of satisfaction.

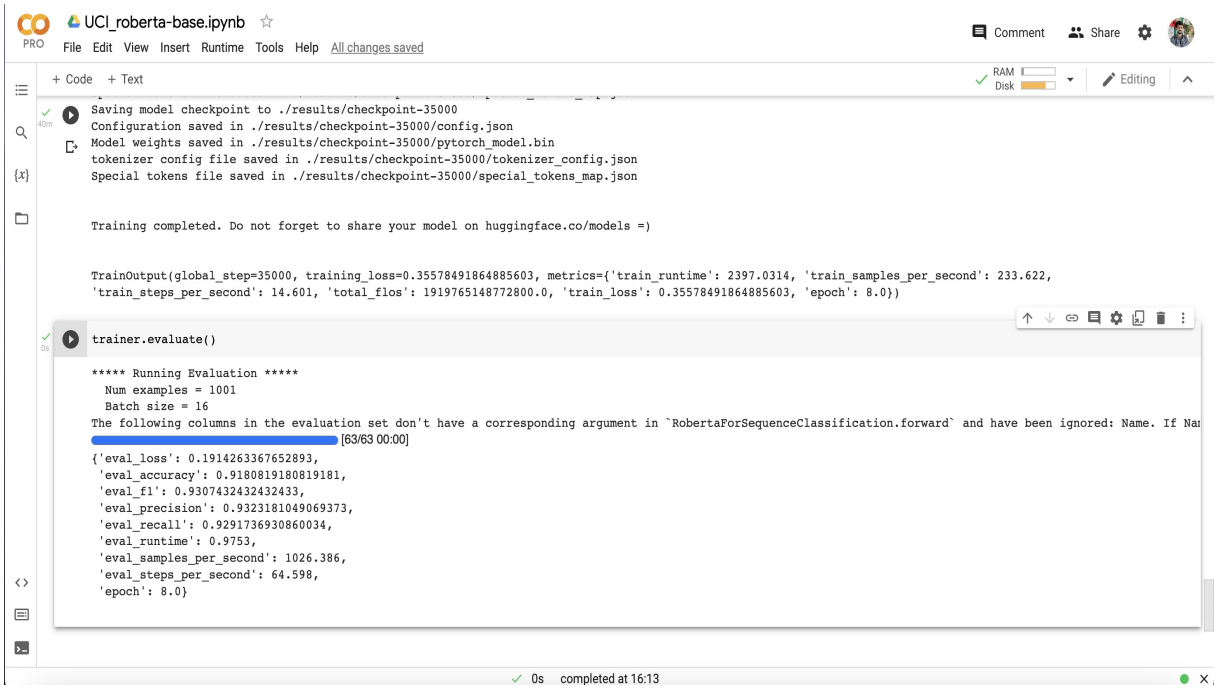

Figure 4.4: Evaluation Results for the RoBERTa model

4.5 Results and Discussion

The table below clearly shows the results of the different models that have been implemented from transformers. These four models are DistilBERT, BERT, XL Net and RoBERTa. BERT model outperformed with 95.90 per cent of accuracy and with the least evaluation loss of 0.08 which make it different from other. Secondly, DistilBERT showed almost the same performance with 95.40 per cent of accuracy and an evaluation loss was 0.09. Both accuracy and evaluation loss has a negligible difference of 0.40 and 0.01 but the major difference between them was execution time. DistilBERT took 19 minutes for the execution whereas BERT model consumed 38 mins for the same. Execution time makes the DistilBERT model more efficient than BERT as it is stated by authors in related work. XL net and RoBERTa also performed well but the accuracy for XL Net was least among all i.e. 88.21 per cent and 91.80 per cent for RoBERTa. Even the execution time of XL Net was more than 64 minutes

[7] https://nlp.johnsnowlabs.com/2021/05/20/roberta_base_en.html

which made it the least efficient model. Time execution for BERT and RoBERTa was almost the same but the accuracy difference was more than 5 per cent. Overall, considering all measures of compute metrics, DistilBERT was the most efficient, fastest and most accurate for this research.

Model	**DistilBERT**	**BERT**	**XL Net**	**RoBERTa**
Evaluation loss (Test Set)	**0.09**	0.08	0.26	0.19
Accuracy	**95.4**	**95.9**	88.21	91.8
F1 Score	**96.09**	96.52	90.05	93.07
Precision	**96.75**	96.93	90.05	93.23
Recall	**95.44**	96.12	90.05	92.91
Epochs	**8**	8	8	8
Execution Time(mins)	**19:04**	38:54	64:57	39:54

Table 4.1: Comparison Table of Results

Conclusion and Future Work

In this research, it can explicitly be observed that DistilBERT and BERT could be used for the gender classification of names. Evaluation of four Fine Tuned Models BERT, DistilBERT, XLNet and RoBERT has been used for the identification of Gender by name. DistilBERT has been stated as the most efficient and fastest model for this work. Tensorflow is also an alternative to reproduce and compare the results. But here, PyTorch has been used for the ultimate finetuning of resources. The research question is precisely answered in this paper that sequence classification using fine-tuned models can classify the gender of the names. All the objectives of the paper are fulfilled and addressed the problems. Overall, it can be concluded that the DistilBERT and BERT models performed and classified the gender with more than expected results. The outcome of all four fine-tuned Models is different, accessed considering "time" as a parameter. Additionally, the misclassified names can be identified and predicted after the extraction of features to classify **unisex** namesn using finetuned models. The findings are that DistilBERT is the fastest and most efficient model lashing all others. For future work reference, in collaboration of these models with other fine-tuned models and a larger dataset can be considered as a large dataset which means larger training steps with lead to better accuracy.

Acknowledgement

I am delighted to pay thanks to my Supervisor, **Mr Prashanth Nayak**, for his able guidance, feedback, effort and time. I am thankful for his valuable advice and continuous support until the end of the research project.

Bibliography

Alam, T., Khan, A. & Alam, F. (2020), 'Bangla text classification using transformers', *arXiv preprint arXiv:2011.04446* .

Ali, D., Missen, M. M. S., Akhtar, N., Salamat, N., Asmat, H. & Firdous, A. (2016), 'Gender prediction for expert finding task', *International Journal of Advanced Computer Science and Applications* **7**(5).

Ani, J. F., Islam, M., Ria, N. J., Akter, S. & Masum, A. K. M. (2021), Estimating gender based on bengali conventional full name with various machine learning techniques, *in* '2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)', IEEE, pp. 1–6.

Çoban, Ö., İnan, A. & Özel, S. A. (2021), 'Facebook tells me your gender: An exploratory study of gender prediction for turkish facebook users', *Transactions on Asian and Low-Resource Language Information Processing* **20**(4), 1–38.

Dolci, T. (2022), Fine-tuning language models to mitigate gender bias in sentence encoders, *in* '2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)', IEEE, pp. 175–176.

Dua, D. & Graff, C. (2017), 'UCI machine learning repository'.
URL: *http://archive.ics.uci.edu/ml*

Grave, E., Bojanowski, P., Gupta, P., Joulin, A. & Mikolov, T. (2018), 'Learning word vectors for 157 languages', *arXiv preprint arXiv:1802.06893* .

Ho Huong, T., Tran-Trung, K. & Truong Hoang, V. (2022), 'A computational linguistic approach for gender prediction based on vietnamese names', *Mobile Information Systems* **2022**.

Jia, J. & Zhao, Q. (2019), Gender prediction based on chinese name, *in* 'CCF International Conference on Natural Language Processing and Chinese Computing', Springer, pp. 676–683.

Mahendra, R., Putra, H. S., Faisal, D. R. & Rizki, F. (2022), 'Gender prediction of indonesian twitter users using tweet and profile features', *Jurnal Ilmu Komputer dan Informasi* **15**(2), 131–141.

Qasim, R., Bangyal, W. H., Alqarni, M. A. & Ali Almazroi, A. (2022), 'A fine-tuned bert-based transfer learning approach for text classification', *Journal of healthcare engineering* **2022**.

Rego, R. C. & Silva, V. M. (2021), 'Predicting gender of brazilian names using deep learning', *arXiv preprint arXiv:2106.10156* .

Shrestha, P., Rey-Villamizar, N., Sadeque, F., Pedersen, T., Bethard, S. & Solorio, T. (2016), Age and gender prediction on health forum data, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 3394–3401.

Soldevilla, I. & Flores, N. (2021), Natural language processing through bert for identifying gender-based violence messages on social media, *in* '2021 IEEE International Conference on Information Communication and Software Engineering (ICICSE)', IEEE, pp. 204–208.

Sun, X. & Huo, X. (2022), 'Word-level and pinyin-level based chinese short text classification', *IEEE Access* .

To, H. Q., Nguyen, K. V., Nguyen, N. L.-T. & Nguyen, A. G.-T. (2020), Gender prediction based on vietnamese names with machine learning techniques, *in* 'Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval', pp. 55–60.

Tripathi, A. & Faruqui, M. (2011), Gender prediction of indian names, *in* 'IEEE Technology Students' Symposium', IEEE, pp. 137–141.

What is CRISP DM? - Data Science Process Alliance (n.d.), `https://www.datascience-pm.com/crisp-dm-2/`. (Accessed on 12/07/2022).

Xu, Z. (2021), 'Roberta-wwm-ext fine-tuning for chinese text classification', *arXiv preprint arXiv:2103.00492* .

Yao, T., Zhai, Z. & Gao, B. (2020), Text classification model based on fasttext, *in* '2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS)', IEEE, pp. 154–157.