CE807 Assignment 2

Offensive Language Classification using OLID dataset

Submitted By

Reg no: 2201829

Introduction

Objective :

- The main objective of this assignment is to classify offensive language using the OLID dataset.
- To compare the performance of two models used for classification with different evaluation matrices and explain the result with suitable examples from the output files of the two models.
- To find the effect of data size on the performance of the model.

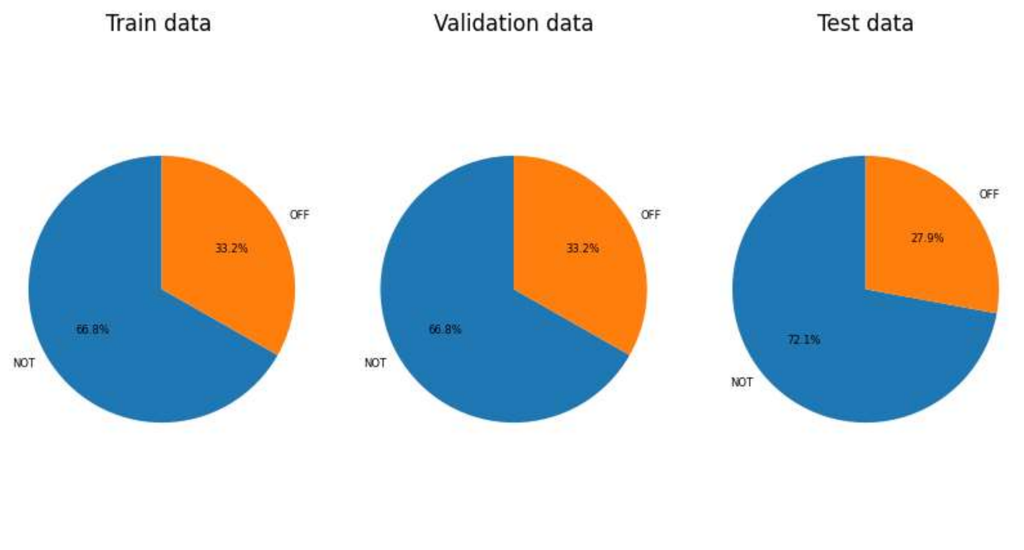
OLID dataset

- OLID – Offensive Language Identification Dataset
- It is publically available over the internet and open for research activities and experiments.
- It is a benchmark dataset used for the classification of offensive language.
- The dataset for this experiment consists of 3 files – train.csv, valid.csv, and test.csv.

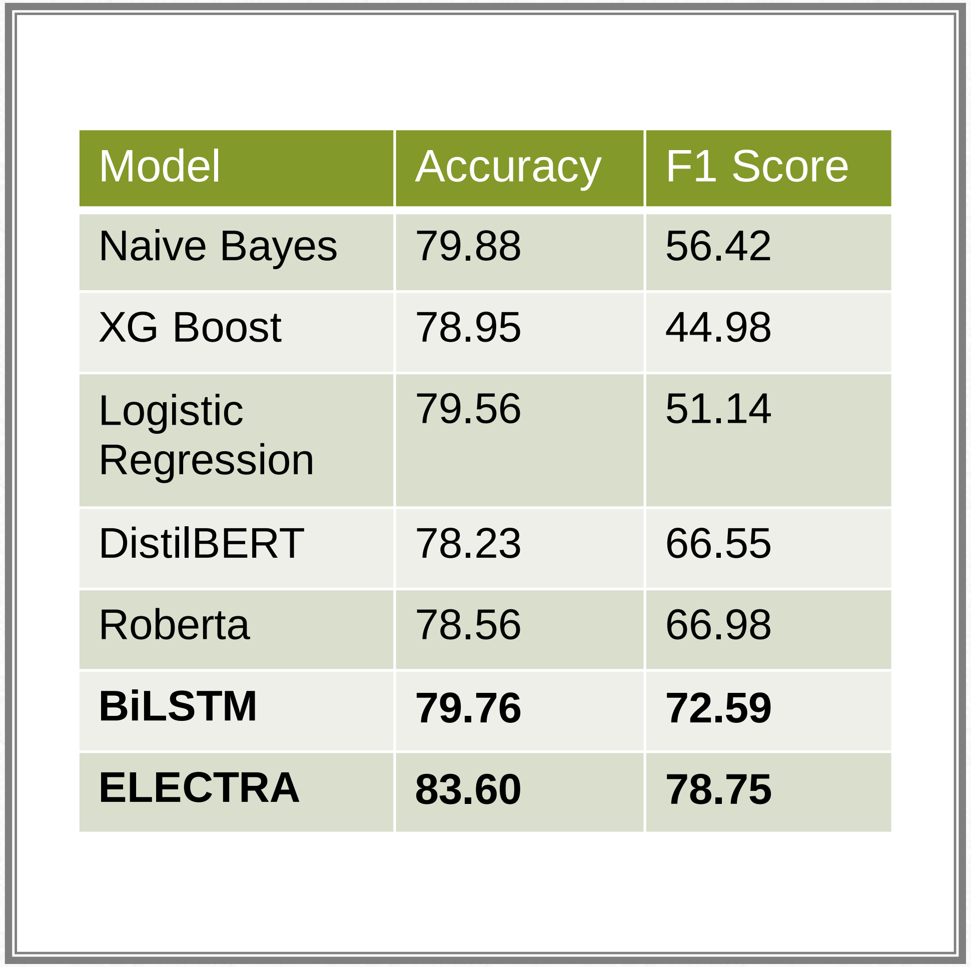
Dataset Exploration

- The following table depicts the total number of data in each data file along with the percentage of data in each class.

Dataset	Total	% OFF	% NOT
Train	12,314	33.23	66.77
Valid	928	33.29	66.81
Test	861	27.99	72.1



Model Selection and Justification



Model	Accuracy	F1 Score
Naive Bayes	79.88	56.42
XG Boost	78.95	44.98
Logistic Regression	79.56	51.14
DistilBERT	78.23	66.55
Roberta	78.56	66.98
BiLSTM	**79.76**	**72.59**
ELECTRA	**83.60**	**78.75**

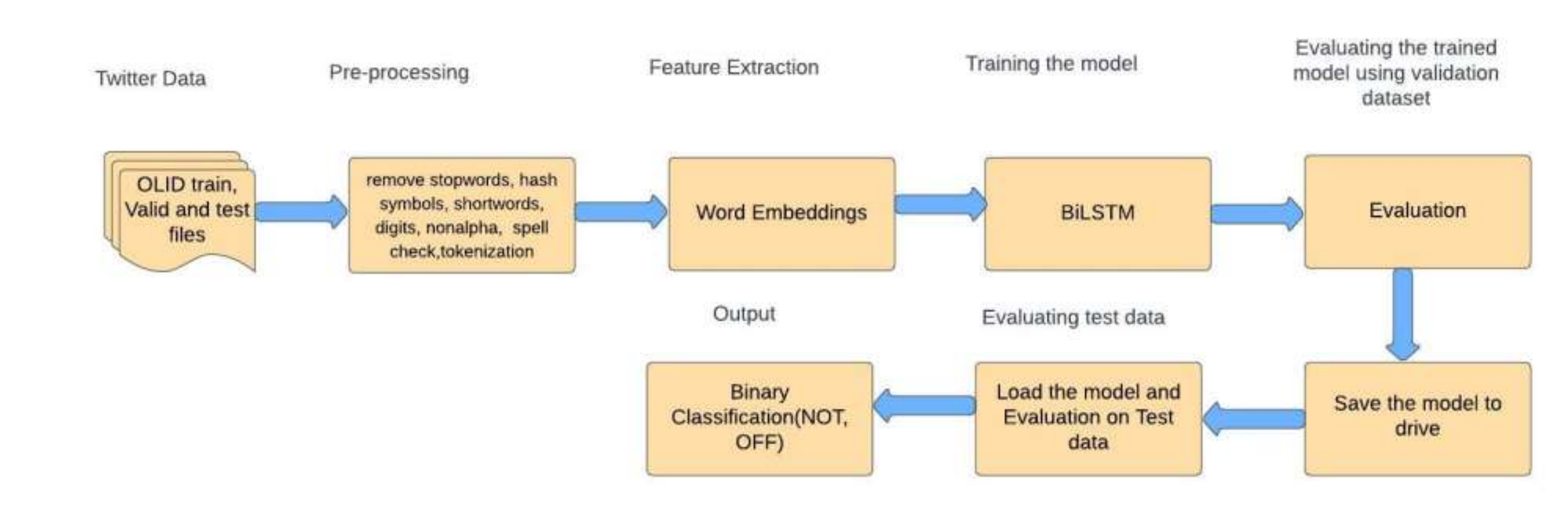
- Have done experiments with 7 models and got state-of-the-art performance for two classifiers and it has been used for this binary text classification task.
- From the table, we can see that for most of the models, the accuracy is between 78-79 % due to the fact that data is unbalanced as the number of 'NOT' classes is approximately 67 in the train and validation set. So, we cannot consider accuracy alone as a matrix for measuring the efficiency and performance of the model.
- While considering the f1-score, it is clear which model performs well, as BiLSTM and ELECTRA were successful in classifying both classes.

Data Preprocessing & Cleaning

This process mainly includes the process of the data and makes it suitable for feeding the model. Following are the data cleaning:

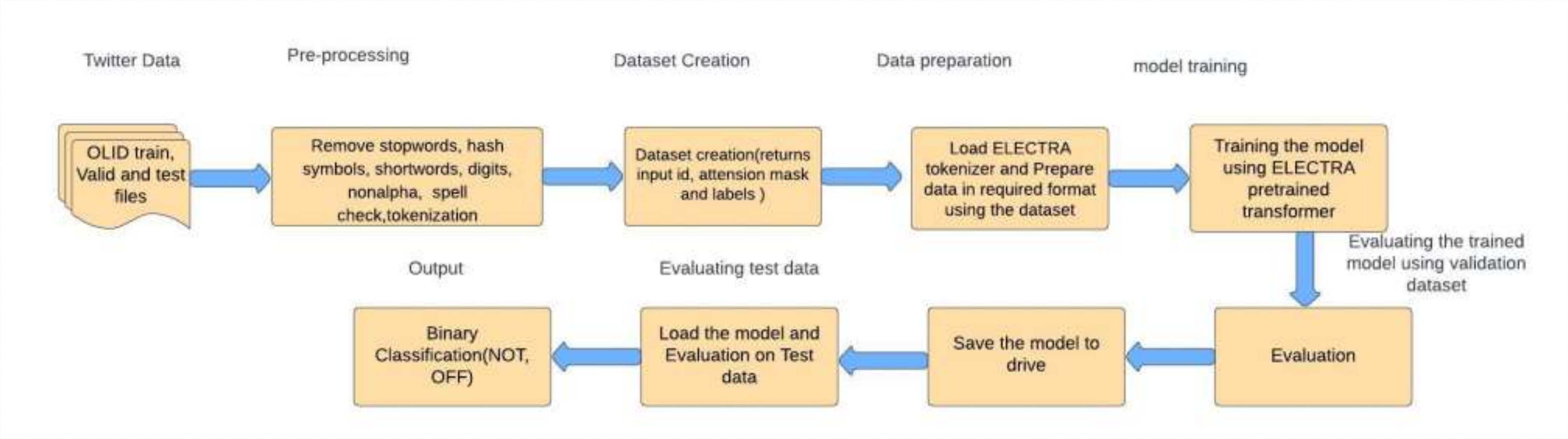
- ❑ Handle the diacritics in the text
- ❑ Removing the stop words
- ❑ Spelling correction
- ❑ Removing the hashtag symbol
- ❑ Removing the short words of length 1 and 2 characters
- ❑ Removing the digits
- ❑ Removing non-alphanumeric values
- ❑ Removing URLs
- ❑ Remove all strings starting with @
- ❑ Removing additional stop words

Pipeline diagram of BiLSTM



Twitter Data
Pre-processing
Feature Extraction
Training the model
Evaluating the trained model using validation dataset
OLID train, Valid and test files
remove stopwords, hash symbols, shortwords, digits, nonalpha, spell check,tokenization
Word Embeddings
BiLSTM
Evaluation
Output
Evaluating test data
Binary Classification(NOT, OFF)
Load the model and Evaluation on Test data
Save the model to drive

Pipeline diagram of ELECTRA



Twitter Data
Pre-processing
Dataset Creation
Data preparation
model training
OLID train, Valid and test files
Remove stopwords, hash symbols, shortwords, digits, nonalpha, spell check,tokenization
Dataset creation(returns input id, attension mask and labels)
Load ELECTRA tokenizer and Prepare data in required format using the dataset
Training the model using ELECTRA pretrained transformer
Evaluating the trained model using validation dataset
Output
Evaluating test data
Binary Classification(NOT, OFF)
Load the model and Evaluation on Test data
Save the model to drive
Evaluation

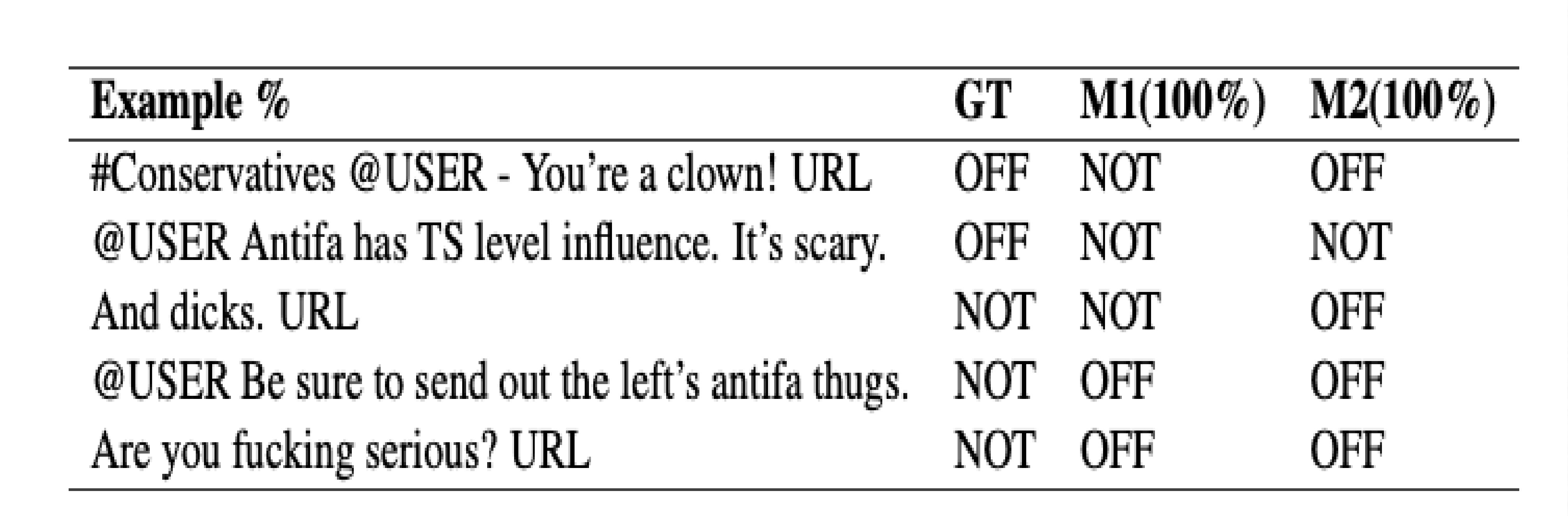
Hyper Parameter Tuning

- Learning Rate
- Maximum input length
- Number of epoch
- Batch size
- Optimizer

Comparison of Model

- From the classification report:
  - ELECTRA outperforms BiLSTM in several performance metrics.
  - ELECTRA has the highest test accuracy of 83.6% compared to BiLSTM with an accuracy of 80.2%, indicating overall better accuracy.
  - ELECTRA has the highest precision score for both classes, indicating its ability to correctly predict positive and negative instances.
  - BiLSTM has a higher recall value for 'NOT', while ELECTRA has a higher recall value for 'OFF', indicating better capability to capture positive instances.
  - ELECTRA also has higher F1 scores for both classes, showing a better balance between precision and recall.
- Based on the evaluation metrics:
  - ELECTRA performs better than BiLSTM in terms of accuracy, F1-score, precision, and recall.
  - This suggests that ELECTRA is better suited for this classification task.

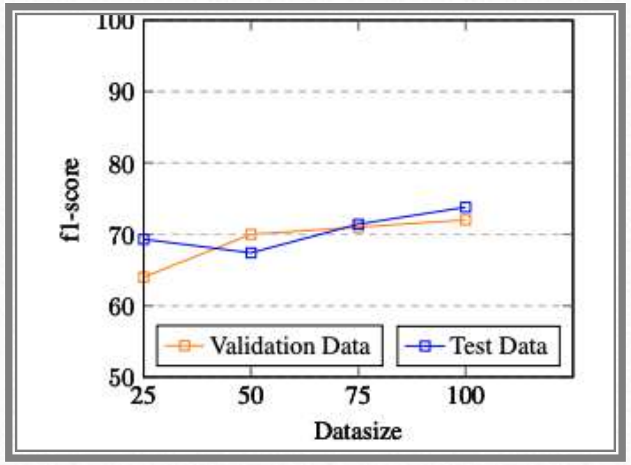
Example of Model Comparison

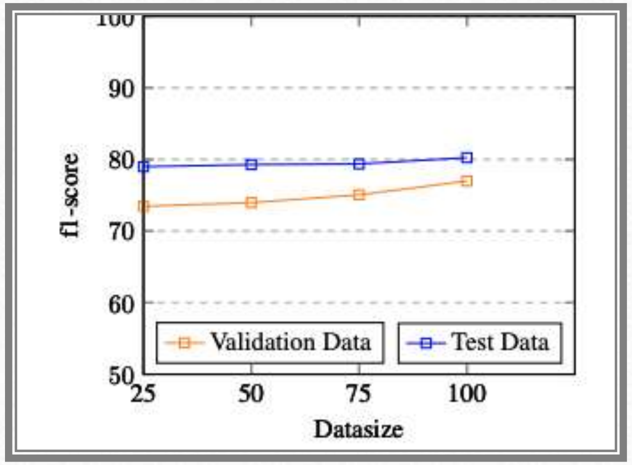


Example %	**GT**	**M1(100%)**	**M2(100%)**
#Conservatives @USER - You're a clown! URL	OFF	NOT	OFF
@USER Antifa has TS level influence. It's scary.	OFF	NOT	NOT
And dicks. URL	NOT	NOT	OFF
@USER Be sure to send out the left's antifa thugs.	NOT	OFF	OFF
Are you fucking serious? URL	NOT	OFF	OFF

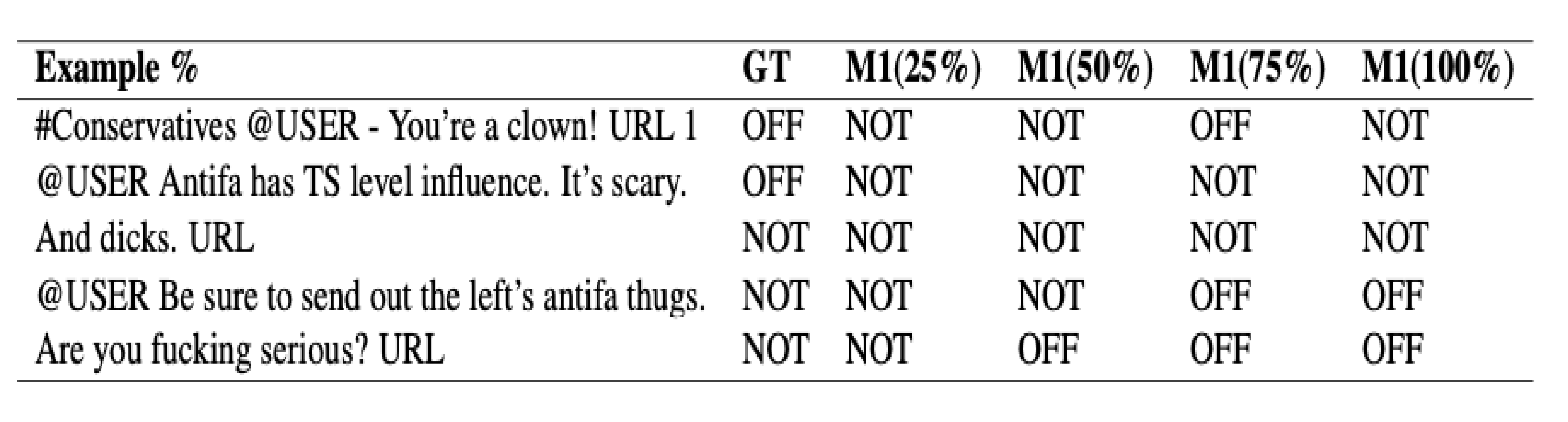
Data Size Comparison

- Datasize comparison has been done with two models by dividing the whole dataset into 25%, 50%, 75% and 100%.
- From the result, it has been identified that the BiLSTM model has an impact on the increasing data set size. While the ELECTRA model consistently performed on the test file will have nearly the same f1 score.
- While in the validation dataset, there is a minor difference of 7% difference in the f1 score of 25% and 100%. Overall we can conclude that ELECTRA can work well with small datasets efficiently.





DataSize Comparison Example



Example %	GT	M1(25%)	M1(50%)	M1(75%)	M1(100%)
#Conservatives @USER - You're a clown! URL 1	OFF	NOT	NOT	OFF	NOT
@USER Antifa has TS level influence. It's scary.	OFF	NOT	NOT	NOT	NOT
And dicks. URL	NOT	NOT	NOT	NOT	NOT
@USER Be sure to send out the left's antifa thugs.	NOT	NOT	NOT	OFF	OFF
Are you fucking serious? URL	NOT	NOT	OFF	OFF	OFF

Conclusion

- The study has been done for implementing a classification model for detecting offensive language using two state-of-the-art models BiLSTM and ELECTRA.
- Both of them performed well with high accuracy, but ELECTRA outperform BiLSTM in all evaluation matrices.
- Also done a comparative study to find the impact of data size on the performance of a model trained with 25%, 50%, 75% and 100% data. And find that BiLSTM has a direct impact on data size while ELECTRA has not had much impact and it almost consistently performed well for all the data sizes.