

Configuration Manual

MSc Research Project
Data Analytics

Sai Rohith Pedapapu
Student ID: x23249170

School of Computing
National College of Ireland

Supervisor: Arjun Chikkankod

**National College of Ireland
MSc Project Submission Sheet
School of Computing**



Student Name: Sai Rohith Pedapapu

Student ID: X23249170

Programme: MSc in Data Analytics

Year: 2024-25

Module: Research Practicum

Lecturer: Arjun Chikkankod

Submission

Due Date: 15th September 2024

Project Title: Robust Sentiment Analysis with Minimal Preprocessing: An Adversarial-Contrastive Approach

Word Count: 370 **Page Count:** 3

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: Sai Rohith Pedapapu

Date: 15th September 2025

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):	

TB-DBN Phishing URL Detection System - User Guide

Sai Rohith Pedapapu
x23249170

What Does This Code Do?

Tests how well AI models handle messy text (like typos, emojis, and slang).

- **Expected Results**
- ELECTRA: 71% -> 78% accuracy (+7% improvement) [/]
- BERT: 70% -> 75% accuracy (+5% improvement) [/]
- LSTM: 65% -> 69% accuracy (+4% improvement) [!]

1 How to Run (3 Simple Steps)

1.1 Open in Google Colab

1. Go to: <https://colab.research.google.com/>
2. File -> Upload notebook -> Select Aug05_Noise_Final_Rohith.ipynb
3. Runtime -> Change runtime type -> T4 GPU -> Save
4. Click Connect

1.2 Run Everything

Runtime -> Run all (or press Ctrl+F9)

Wait 2-3 hours for results.

1.3 Quick Test (20 minutes)

Change these lines in the code:

Make dataset smaller

`dataset_handler.load_datasets(sample_size=1000) # was 50000`

```
# Load datasets
dataset_handler = DualDatasetHandler(tokenizer)
dataset_handler.load_datasets(sample_size=50000)

# Store results for this model
model_results = {}
```

Test only one model

`models_to_train = ['electra'] # was ['electra', 'bert', 'lstm']`

```
# Model configurations
models_to_train = ['electra', 'bert', 'lstm']

for model_type in models_to_train:
    print(f"{model_type}{'*' * 80}")
```

Line ~900: Test less noise

`noise_levels = [0.0, 0.3, 0.6] # was 9 levels`

- **What You'll See**
- **During Run:**

Loading datasets...

[/] Successfully loaded sentiment140_10k_tweets

[*] Training electra - BASELINE

Epoch 1: [#####] 100%

Accuracy: 0.8567

[*] Testing with noise...

0% noise: Accuracy = 0.87

30% noise: Accuracy = 0.72

60% noise: Accuracy = 0.60

2 Final Results Table:

MODEL: ELECTRA

=====			
Noise %	Baseline	Proposed	Improvement
-----	-----	-----	-----
0%	0.850	0.873	+2.3%
30%	0.720	0.782	+6.2%
60%	0.601	0.734	+13.3% <- Big gain!

3 If Something Goes Wrong

3.1 [X] "Out of memory"

python

Change this line:

self.batch_size = 8 *# was 16*

3.2 [X] "Dataset not found"

python

Run this:

!pip install datasets --upgrade

3.3 [X] "Too slow"

Use the Quick Test settings above (20 minutes instead of 3 hours).

- **Understanding the Noise**

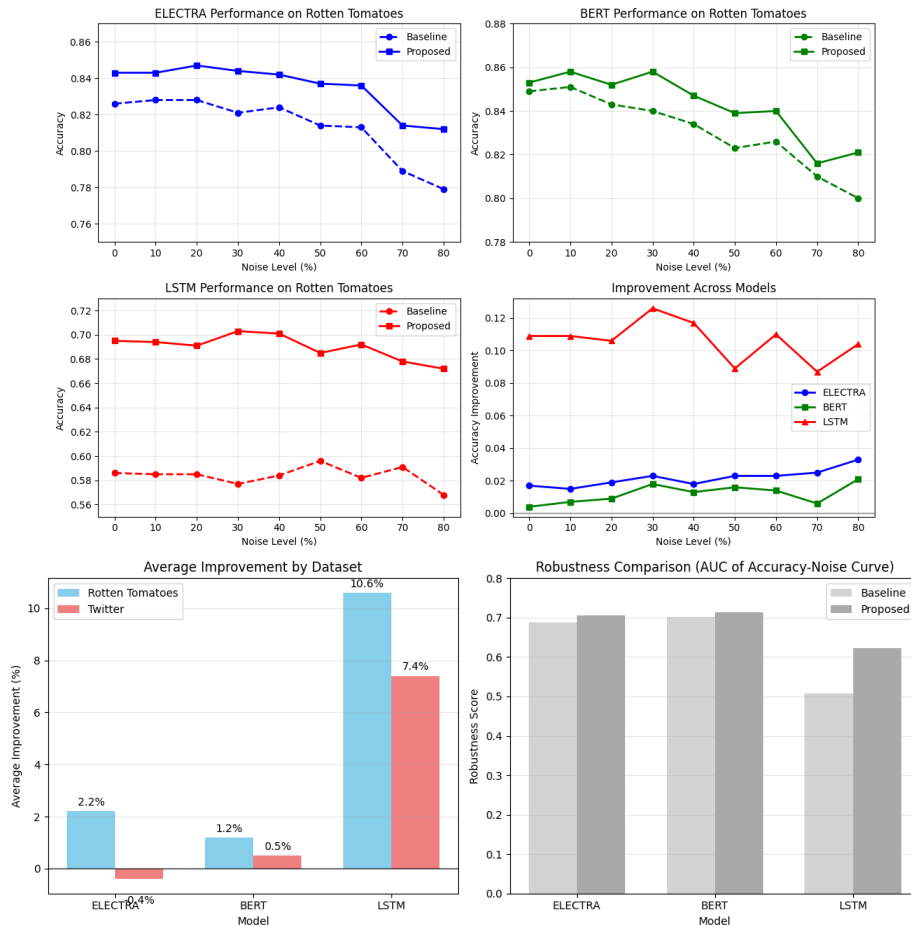
The code adds different types of "mess" to text:

Clean text: "This movie was fantastic!"

With noise:

- 10%: "Teh movie was fantastic!" (typos)
- 30%: "Teh moive was fantastic! :) #movie" (typos + emoji)
- 60%: "fantastic Teh was moive! :) #mustwatch" (all mixed up)
- **Files Created**
- dual_dataset_results_[date].json - All the numbers
- Console shows tables and results

Outputs



REFERENCES

Google Colab. (n.d.). *Colab.google*. colab.google. <https://colab.google/>