TRADITIONAL METHODS FOR HINDI NAMED ENTITY RECOGNTION

TEAM MEMBERS:
Akshara Choutapally(2020IMT-024)
Deepjyoti Boro(2020IMT-025)
Devanshu Patidar(2020IMT-026)

Presented to Dr. Sunil Kumar

TABLE OF CONTENT

Named Entity Recognition

- NER is a vital NLP task used to **extract and categorize named entities into predefined classes** such as person, location, organization, numeral, and temporal entities.[1]

- **Some Applications of NER:**

  Information Retrieval
  Question Answering
  Sentimental Analysis

- Enables better understanding of text context and semantics.

Cristiano Ronaldo PERSON dos Santos Aveiro PERSON GOIH ComM (Portuguese pronunciation: [krɪʃˈtjɐnu ʁɔˈnaɫdu PERSON]; born 5 February 1985 DATE) is a Portuguese professional footballer who plays as a forward for Premier League club Manchester United and captains the Portugal national team. He has won 32 trophies in his career, including seven league titles, five UEFA Champions Leagues, one UEFA European Championship EVENT , and one UEFA Nations League. Ronaldo PERSON holds the records for most appearances (183), most goals (140), and assists (42) in the Champions League, most goals in the European Championship EVENT (14), most international goals by a male player (115), and most international appearances by a European male (186).

SOME GLIMPSE OF HINER DATASET

```
datasets
```

Python

```
DatasetDict({
    train: Dataset({
        features: ['id', 'tokens', 'ner_tags'],
        num_rows: 75827
    })
    validation: Dataset({
        features: ['id', 'tokens', 'ner_tags'],
        num_rows: 10851
    })
    test: Dataset({
        features: ['id', 'tokens', 'ner_tags'],
        num_rows: 21657
    })
})
```

SOME GLIMPSE OF HINER DATASET

```python
datasets["train"][2]
```

Python

```
{'id': '2',
 'tokens': ['रामनगर',
  'इगलास',
  ',',
  'अलीगढ़',
  ',',
  'उत्तर',
  'प्रदेश',
  'स्थित',
  'एक',
  'गाँव',
  'है।'],
 'ner_tags': [0, 0, 6, 0, 6, 0, 3, 6, 6, 6, 6]}
```

SOME GLIMPSE OF HINER DATASET

```python
label_list = datasets["train"].features[f"ner_tags"].feature.names
label_list
```

Python

```
['O', 'B-PER', 'I-PER', 'B-LOC', 'I-LOC', 'B-ORG', 'I-ORG']
```

'O': Outside of any named entity

'B-PER': Beginning of a person's name

'I-PER': Inside of a person's name (if it consists of multiple tokens)

'B-LOC': Beginning of a location name

'I-LOC': Inside of a location name (if it consists of multiple tokens)

'B-ORG': Beginning of an organization name

'I-ORG': Inside of an organization name (if it consists of multiple tokens)

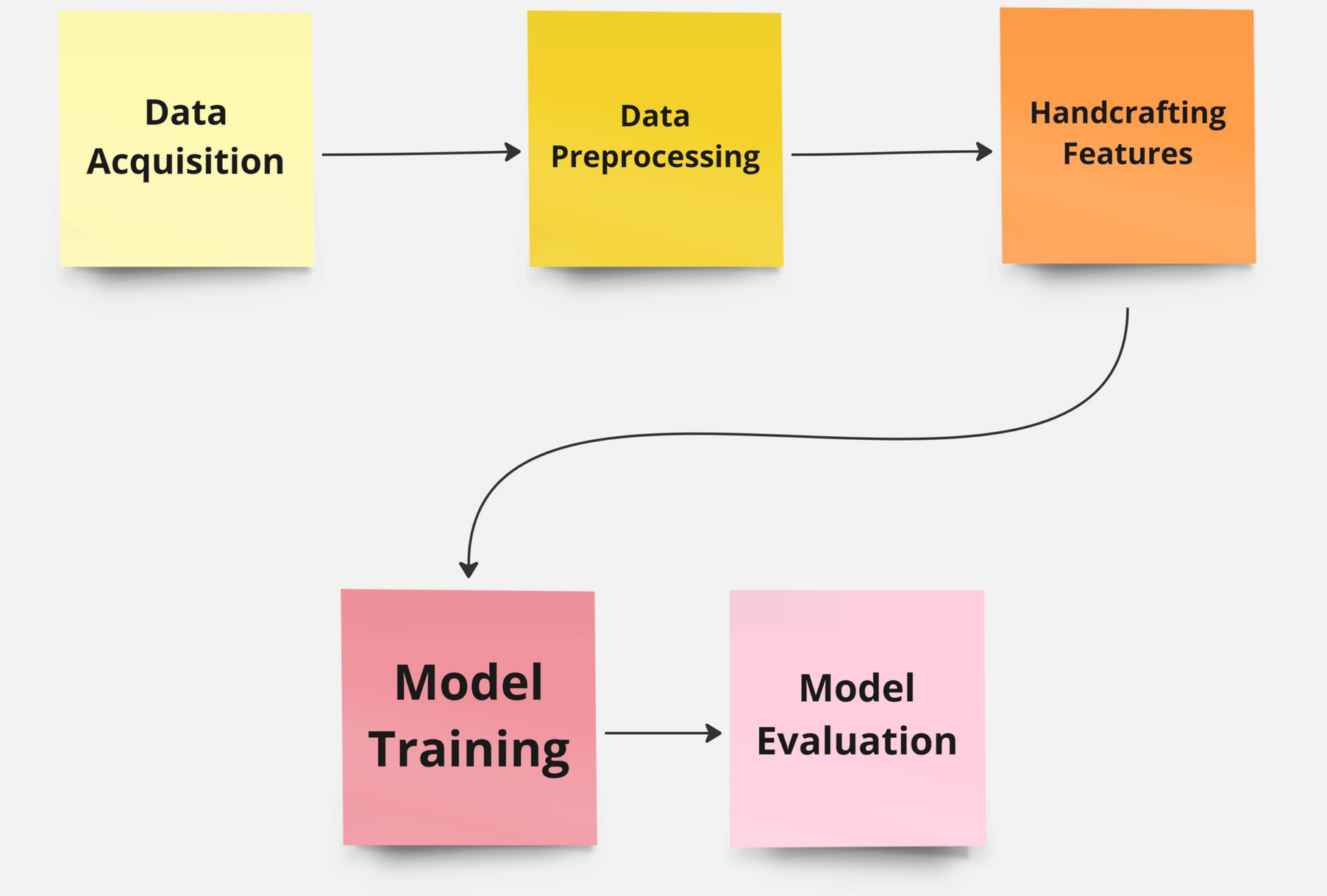
SOME GLIMPSE OF HINER DATASET

```
show_random_elements(datasets["train"])
```

Python

	id	tokens	ner_tags
0	61733	[जल्दी, ही, रामनरेश, सरवन, भी, 5, रन, बना, कर, पैवेलियन, लौट, गए, .जब, सरवन, के, रूप, में, पाँचवां, विकेट, गिरा, वेस्टइंडीज़, के, 12वें, ओवर, में, मात्र, 64, रन, बने, थे, .]	[I-ORG, I-ORG, I-PER, B-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, O, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]
1	59780	[मेरा, साथी, 1985, में, बनी, हिन्दी, भाषा, की, फिल्म, है।]	[I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]
2	59142	[१४४, किलोमीटर, लंबा, यह, राजमार्ग, थंजावुर, को, मनमदुरई, से, जोड़ता, है।]	[I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, O, I-ORG, O, I-ORG, I-ORG, I-ORG]
3	65610	[जब, बर्नार्ड, मैडॉफ़, ने, सभी, 11, आरोप, स्वीकार, कर, लिए, तो, न्यायाधीश, ने, इसे, मंज़ूरी, दे, दी, .]	[I-ORG, I-PER, B-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]
4	75660	[ज़ोया, अफ़रोज़, (जन्म:, १०, जनवरी, १९९४,), एक, भारतीय, अभिनेत्री, और, मॉडल, हैं।]	[I-PER, B-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG, I-ORG]
5	17302	[इंतज़ाम, .]	[I-ORG, I-ORG]

METHODOLOGY


Data Acquisition
Data Preprocessing
Handcrafting Features
Model Training
Model Evaluation

DATA ACQUISITION

- **HiNER dataset** available from the **`datasets` library** was used.[2]
- This dataset is specifically designed for NER tasks in Hindi and includes a variety of annotated named entities.
- The dataset was divided into a training set and a test set, with the training set used to train the models and the test set used to evaluate their performance.

DATA PREPROCESSING

- Data preprocessing is crucial to prepare the raw data for effective machine learning. The preprocessing steps included:

- **Tokenization:** Splitting the text data into individual words or tokens. Tokenization is essential for creating features from the text data.

- **Part-of-Speech (POS) Tagging:** Each token in the dataset was annotated with a part-of-speech tag **using the StanfordNLP library** configured for Hindi. POS tags provide syntactic information about each word, which is valuable for understanding its role in a sentence. [3]

- **Data Limiting:** Due to computational constraints or to balance the dataset, the number of sentences used from the dataset was limited to the **first 2000 entries for both training and testing phases.**

FEATURE ENGINEERING

- Feature engineering is a critical step that involves creating input features for the machine learning models from the raw data. The features extracted for this project included:

- **POS Tags:** The POS tag of each word was used as a feature because different types of words (nouns, verbs, adjectives) tend to have different likelihoods of being named entities.

- **Binary Indicators:** Features indicating whether a word is at the beginning **(`BOS`) or end (`EOS`)** of a sentence were included. These features help the model recognize boundary-related patterns that are common in named entities**.**

- **Contextual Features:** The **words immediately before and after a given word were used as features**. It's important in NER tasks as the likelihood of a word being a named entity can depend heavily on its surrounding words.

MODEL TRAINING AND EVALUATION

- With the features prepared, the next step was to train different machine learning models. Three models were chosen for this study:[4]
  Decision Tree Classifier
  Random Forest Classifier
  Support Vector Machine(SVM)

- Each model was trained on the extracted features and were tested using the unseen test set.

- The **evaluation metrics** used were **precision, recall, and F1-score**

- **Confusion matrices** were generated for each model to visually assess how well each model was performing across different classes of named entities.

PERFORMANCE EVALUATION

PERFORMANCE METRICS USED

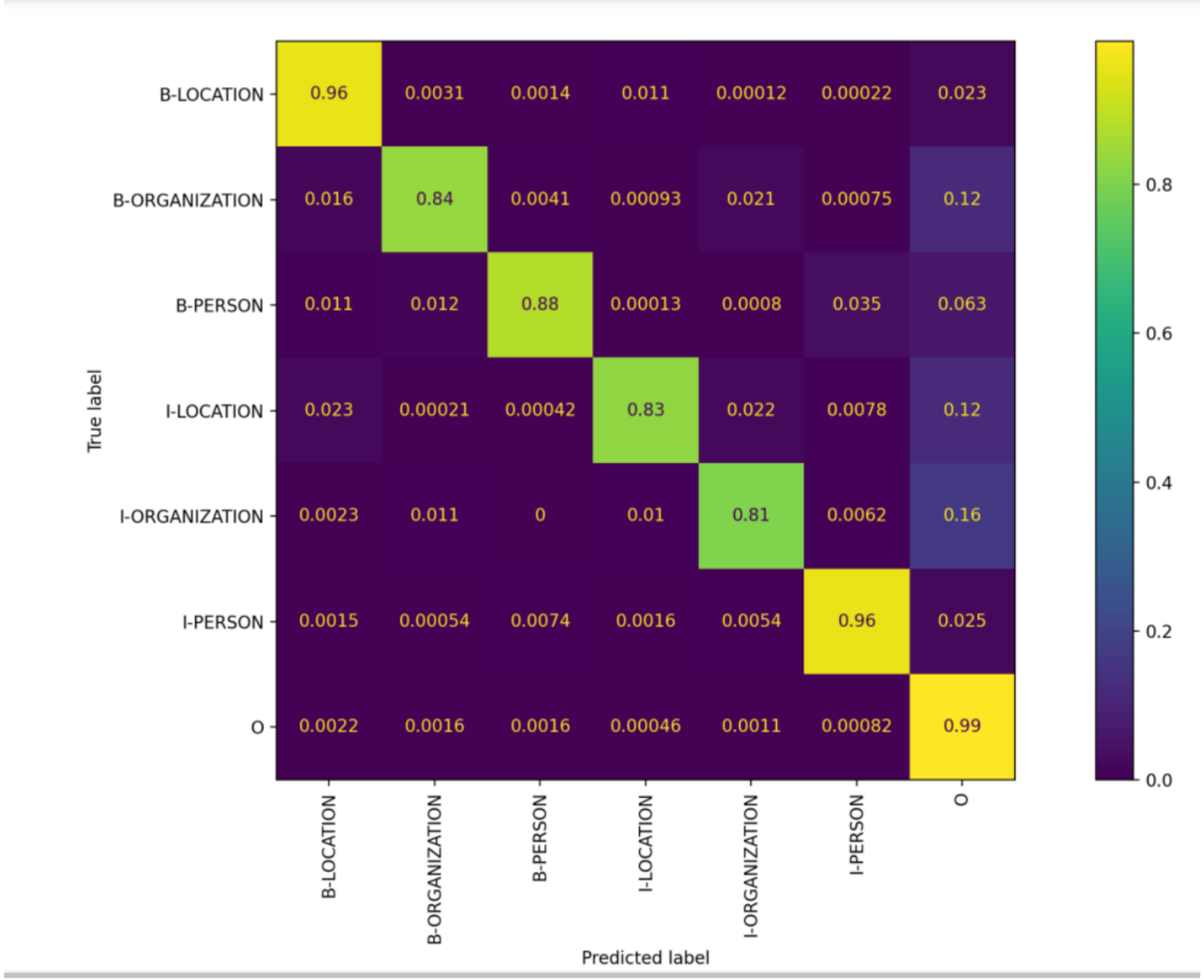
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

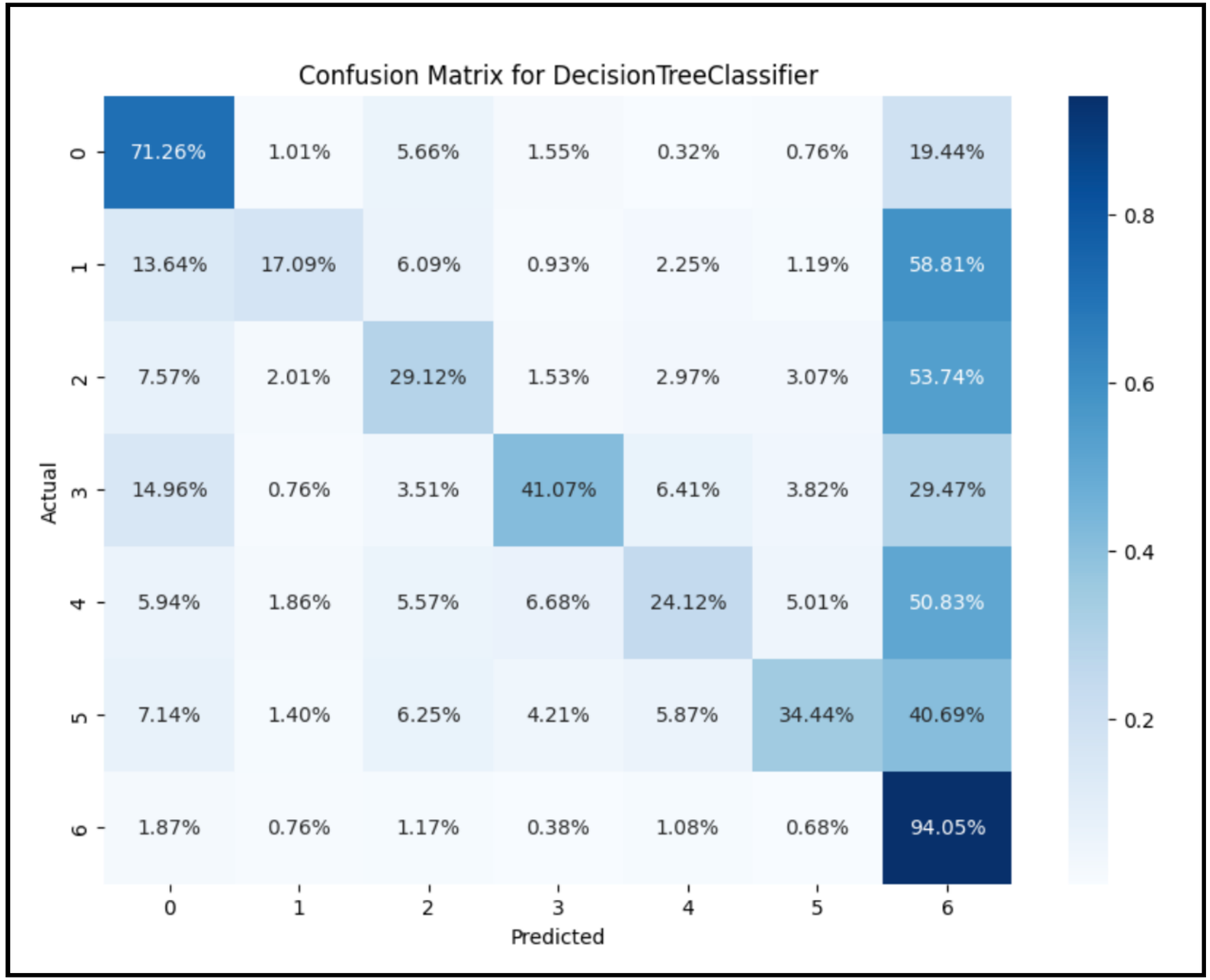
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

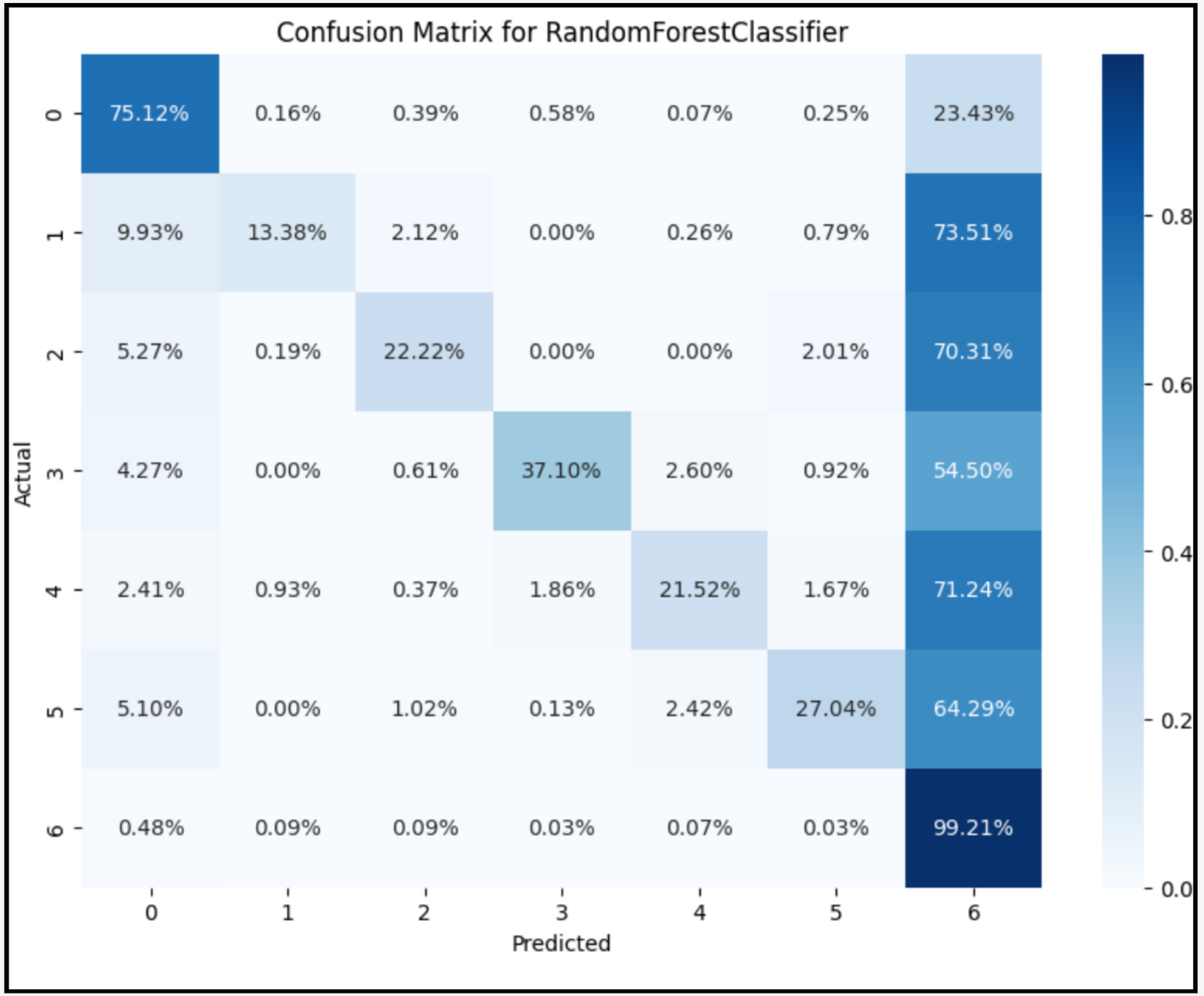
MODEL	Weighted F1 score	Macro Avg F1 score
Base Paper Model	**0.92**	**0.86**
Random Forest	0.91	0.49
Decision Tree	0.89	0.43
SVM	0.89	0.35

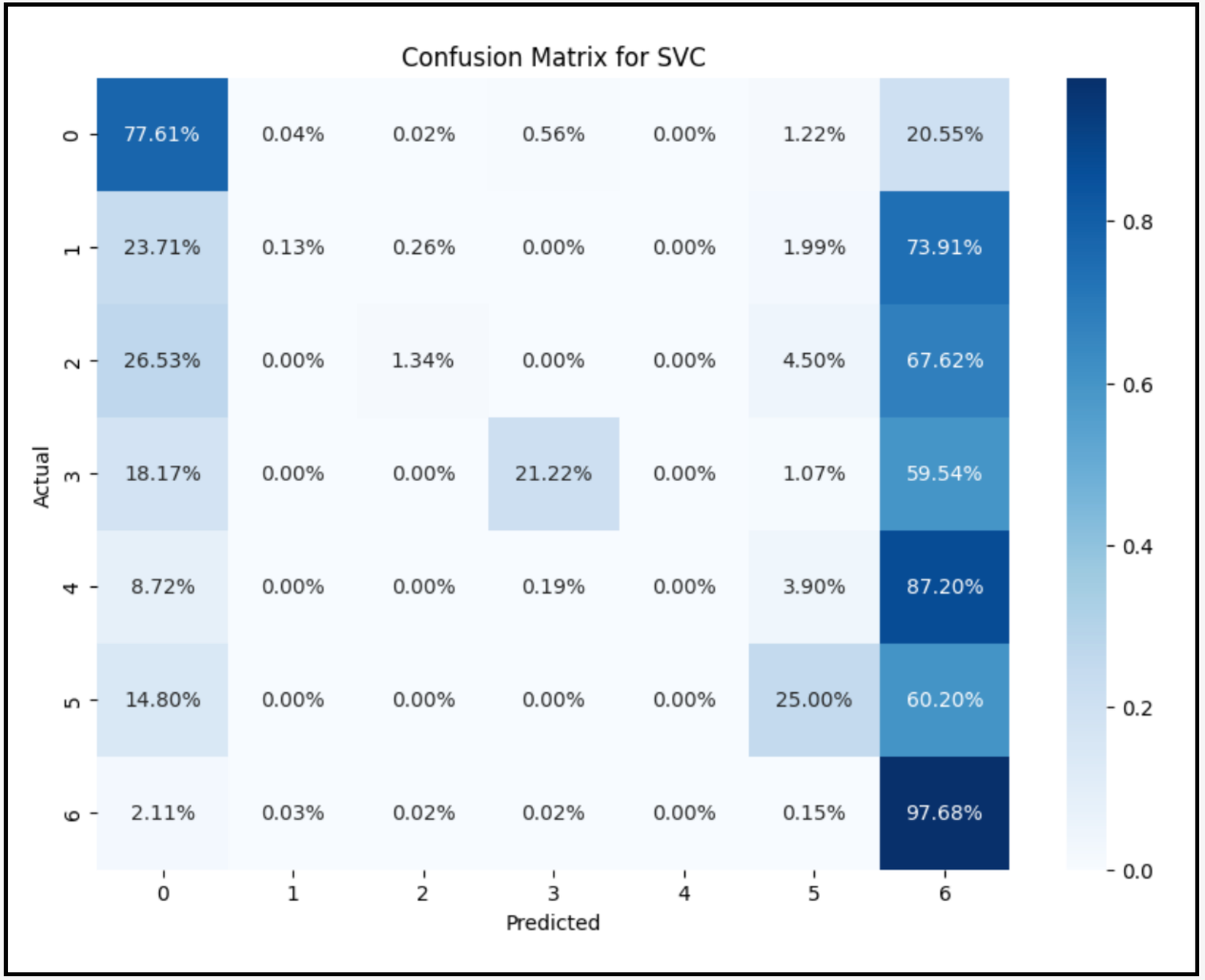
BASE PAPER RESULTS



OUR RESULTS:


Confusion Matrix for DecisionTreeClassifier
Actual
Predicted
0
1
2
3
4
5
6
71.26% 1.01% 5.66% 1.55% 0.32% 0.76% 19.44%
13.64% 17.09% 6.09% 0.93% 2.25% 1.19% 58.81%
7.57% 2.01% 29.12% 1.53% 2.97% 3.07% 53.74%
14.96% 0.76% 3.51% 41.07% 6.41% 3.82% 29.47%
5.94% 1.86% 5.57% 6.68% 24.12% 5.01% 50.83%
7.14% 1.40% 6.25% 4.21% 5.87% 34.44% 40.69%
1.87% 0.76% 1.17% 0.38% 1.08% 0.68% 94.05%
0.8
0.6
0.4
0.2


Confusion Matrix for RandomForestClassifier
Actual
Predicted
0
1
2
3
4
5
6
75.12%
0.16%
0.39%
0.58%
0.07%
0.25%
23.43%
9.93%
13.38%
2.12%
0.00%
0.26%
0.79%
73.51%
5.27%
0.19%
22.22%
0.00%
0.00%
2.01%
70.31%
4.27%
0.00%
0.61%
37.10%
2.60%
0.92%
54.50%
2.41%
0.93%
0.37%
1.86%
21.52%
1.67%
71.24%
5.10%
0.00%
1.02%
0.13%
2.42%
27.04%
64.29%
0.48%
0.09%
0.09%
0.03%
0.07%
0.03%
99.21%
0.8
0.6
0.4
0.2
0.0


Confusion Matrix for SVC
77.61%
0.04%
0.02%
0.56%
0.00%
1.22%
20.55%
23.71%
0.13%
0.26%
0.00%
0.00%
1.99%
73.91%
26.53%
0.00%
1.34%
0.00%
0.00%
4.50%
67.62%
18.17%
0.00%
0.00%
21.22%
0.00%
1.07%
59.54%
8.72%
0.00%
0.00%
0.19%
0.00%
3.90%
87.20%
14.80%
0.00%
0.00%
0.00%
0.00%
25.00%
60.20%
2.11%
0.03%
0.02%
0.02%
0.00%
0.15%
97.68%
Actual
0
1
2
3
4
5
6
Predicted
0.0
0.2
0.4
0.6
0.8

Classification Report for DecisionTreeClassifier:

	precision	recall	f1-score	support
0	0.81	0.71	0.76	3778
1	0.20	0.17	0.19	508
2	0.26	0.24	0.25	671
3	0.35	0.38	0.36	431
4	0.26	0.27	0.26	349
5	0.22	0.31	0.26	495
6	0.95	0.95	0.95	34477
accuracy			0.89	40709
macro avg	0.44	0.43	0.43	40709
weighted avg	0.89	0.89	0.89	40709

Classification Report for RandomForestClassifier:

	precision	recall	f1-score	support
0	0.90	0.74	0.81	3778
1	0.65	0.14	0.23	508
2	0.58	0.20	0.30	671
3	0.82	0.35	0.49	431
4	0.64	0.19	0.29	349
5	0.62	0.23	0.33	495
6	0.93	0.99	0.96	34477
accuracy			0.92	40709
macro avg	0.73	0.41	0.49	40709
weighted avg	0.91	0.92	0.91	40709

Classification Report for SVC:

	precision	recall	f1-score	support
0	0.71	0.78	0.74	3778
1	0.08	0.00	0.00	508
2	0.50	0.02	0.05	671
3	0.92	0.23	0.37	431
4	0.00	0.00	0.00	349
5	0.43	0.25	0.32	495
6	0.93	0.98	0.95	34477
accuracy			0.91	40709
macro avg	0.51	0.32	0.35	40709
weighted avg	0.88	0.91	0.89	40709

CONCLUSION

BASE PAPER MODEL OUTPERFORMED TRADITIONAL METHODS BY LARGE MARGIN.

why?

- **Limited training data** as a **consequence of limited computational resources**.
- **Traditional methods not suitable to handle sequential data** and patterns.
- **Traditional methods rely on hand-crafted features**, while **modern models can automatically learn representations from raw text data.**
- **Lack of tools and library support for regional languages.**

FUTURE SCOPE

- **Use of models** that **capture temporal dependencies** in the text data.
- **Use of Models** that understand **Contextual nature of the tokens**.
- **Improving feature engineering** , adding morphological features.
- **Generating text classification datasets, sentiment analysis datasets** from this dataset.

REFERENCES

[1]A. A. Awan, “What is Named Entity Recognition (NER)?,” Datacamp. https://www.datacamp.com/blog/what-is-named-entity-recognition-ner

[2] R. Murthy, P. Bhattacharjee, R. Sharnagat, J. Khatri, D. Kanojia, and P. Bhattacharyya, “HiNER: A Large Hindi Named Entity Recognition Dataset.” Accessed: Apr. 09, 2024. [Online]. available:https://www.cse.iitb.ac.in/~pb/papers/lrec22-ner.pdf

[3]“stanfordnlp/stanza,” GitHub, https://github.com/stanfordnlp/stanza/

[4]P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, and P. Ghosh, “A Survey on Recent Named Entity Recognition and Relationship Extraction Techniques on Clinical Texts,” Applied Sciences, vol. 11, no. 18, p. 8319, Sep. 2021, doi: https://doi.org/10.3390/app11188319.

THANK YOU