ONLINE CONVERSATIONAL TOXICITY

PRASOON KARMACHARYA

BACKGROUND

Toxic behaviour is pervasive in every online environment

40% internet users have faced harassment

73% internet users have seen others get harassed

26% women aged 18-24 years, have received obscene text

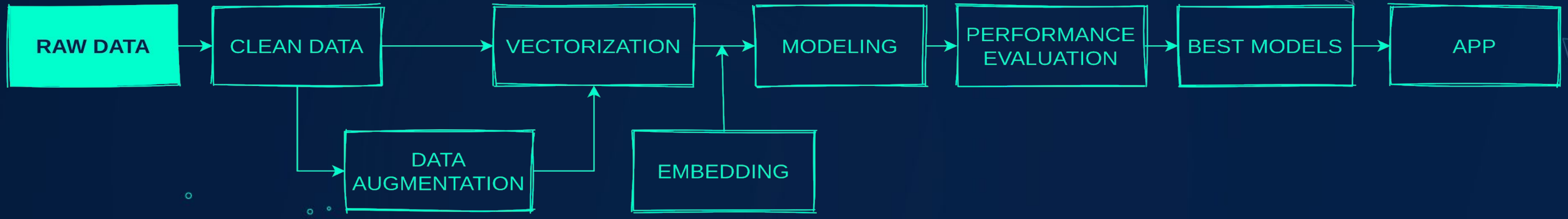
27% internet user chose not to post something online after seeing someone is harassed

PROBLEM STATEMENT

Online communication can often devolve into abuse and harassment due to the anonymity of users. Platforms often struggle to effectively facilitate conversation forcing many communities to shut down user comments. This discourages civil and productive discourse.

Can we leverage natural language processing and machine learning to construct a model that can accurately classify toxicity level in online conversations?

MODELING WORKFLOW



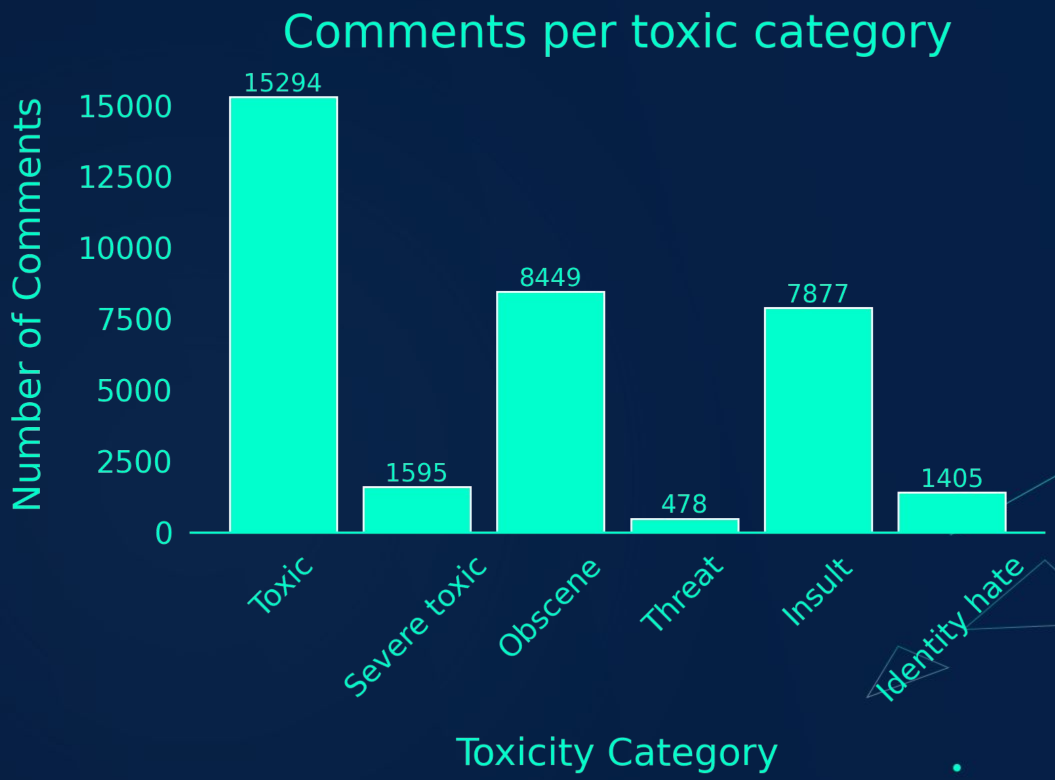
DATA

Source: Civil Comment Corpus (2017) from Jigsaw/Conversation AI [1] and WMF

Data Schema:

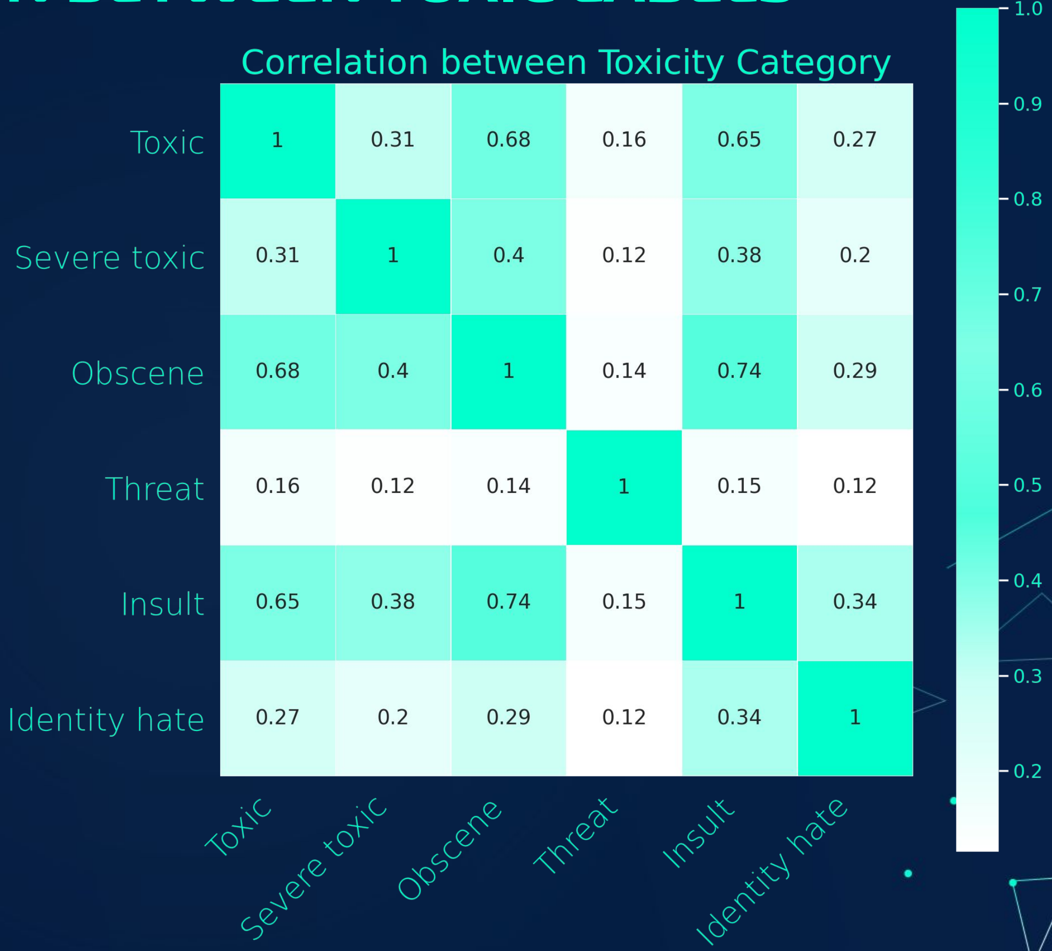
- **Data points:** 159, 571 labeled samples of Wikipedia comments
- **Feature:** Comment Text
- **Binary Labels:** Toxic, Severe Toxic, Obscene, Threat, Insult and Identity Hate

DISTRIBUTION OF TOXIC COMMENTS



Extreme Class Imbalance Problem

CORRELATION BETWEEN TOXIC LABELS



CLASSIFIERS

CLASSIFIERS

DATA

Comment containing at least one toxic class label is considered **neutral** for **binary classification**

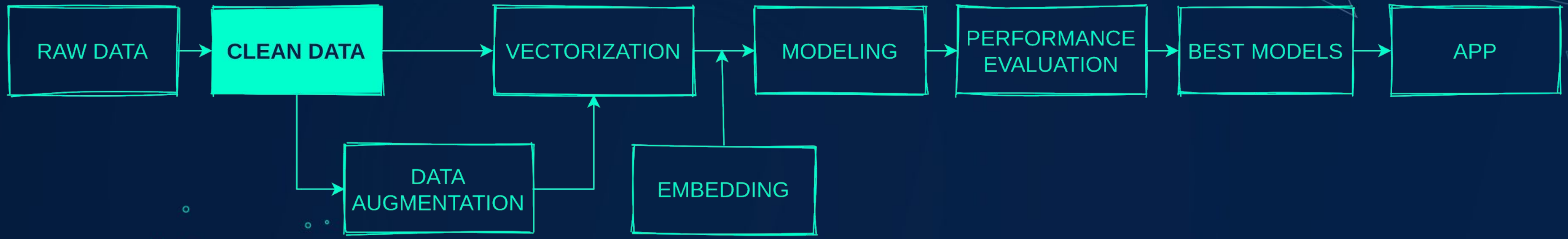
BINARY

Toxic Vs. Neutral

MULTILABEL

Toxicity Classification

MODELING WORKFLOW

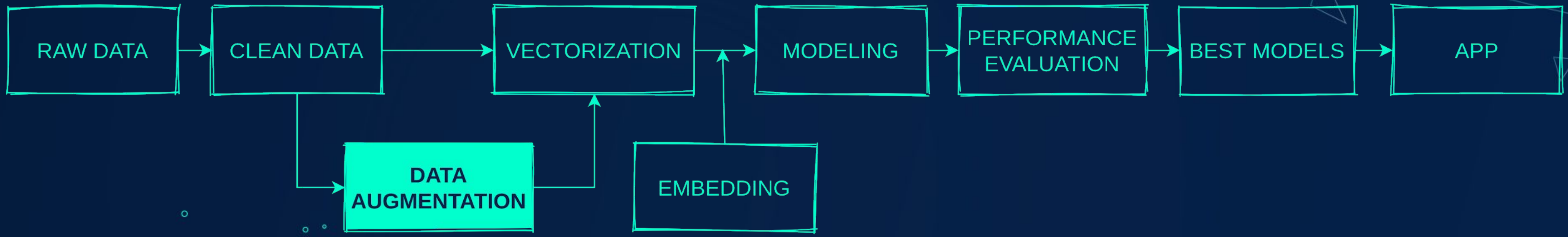


DATA CLEANING

Text cleaning procedure involved:

- Removing emails, urls, dates
- Removing html, xml tags
- Removing non-alphanumeric characters (@#$%^&* etc.)
- Keeping only one instance of copy-pasted comment text

MODELING WORKFLOW



DATA AUGMENTATION

Effort to minimize class-imbalance

Language Translation

Original Text: “Essence of mathematics lies in its freedom.”

Augmented Text: “Abstract mathematics lies in its independence.”

DATA AUGMENTATION

Effort to minimize class-imbalance

Synonym Replacement

Original Text: "The Universe is under no obligation to make sense to you"

Augmented Text: "The Universe is under no obligation to make feel to you"

DATA AUGMENTATION

Effort to minimize class-imbalance

Random Insertion

Original Text: “Premature optimization is root of all evils.”

Augmented Text: “Premature optimization is immorality of all evils.”

DATA AUGMENTATION

Effort to minimize class-imbalance

Random Swap

Original Text: “The future belongs to those who believe in the beauty of their dreams.”

Augmented Text: “The future belongs to those who believe in the dreams of their beauty.”

DATA AUGMENTATION

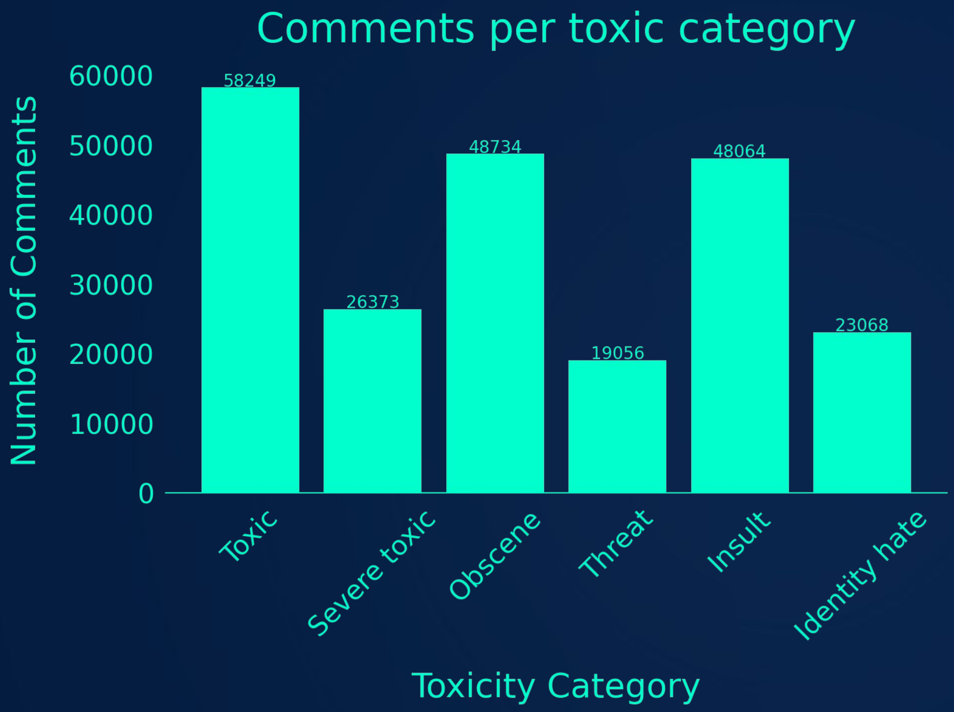
Effort to minimize class-imbalance

Random Deletion

Original Text: “Life is really simple, but we insist on making it complicated.”

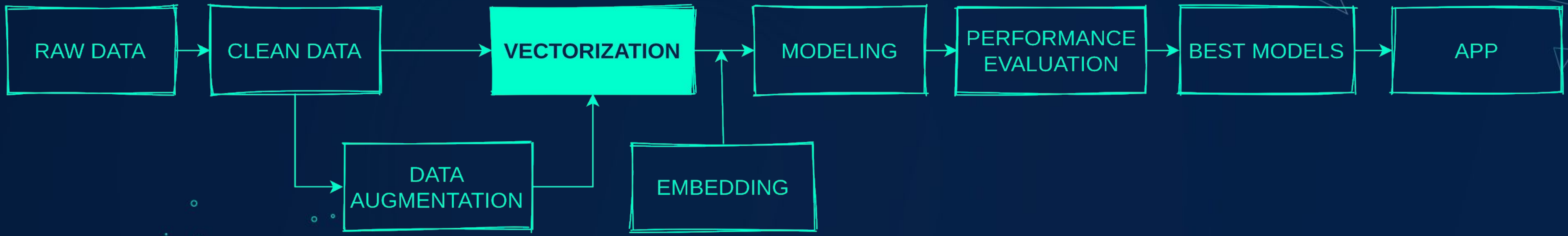
Augmented Text: “Life is really simple, but _ insist _ making _ complicated.”

DATA AUGMENTATION



Minimized Class Imbalance Problem

MODELING WORKFLOW



VECTORIZATION

Tokenization

The process of turning human language text input into numeric data (a numerical representation that the computer can understand.)

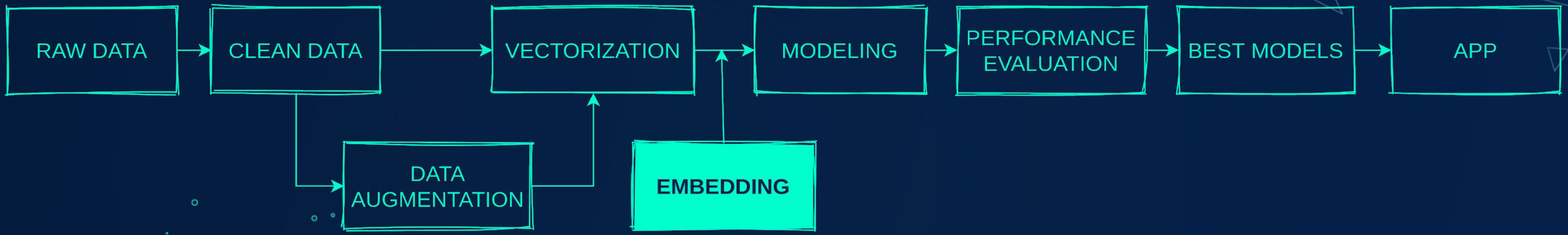
```
sentences = [
    'I love my dog',
    'I love my cat',
    'You love my dog!',
    'Do you think my dog is amazing?'
]
```

```
{'amazing': 10, 'dog': 3, 'you': 5, 'cat': 6,
 'think': 8, 'i': 4, 'is': 9, 'my': 1, 'do': 7,
 'love': 2}

[[4, 2, 1, 3], [4, 2, 1, 6], [5, 2, 1, 3], [7,
5, 8, 1, 3, 9, 10]]
```

Convert Natural Language to Numeric Representation

MODELING WORKFLOW



EMBEDDINGS

Embeddings are pretrained numerical representation of words and phrases in a corpus, that capture their meaning, semantic relationships and sentence morphology.

Pretrained Embeddings

2013

Word2vec

Mikolov et. al. [2]

Word level n-gram model

2014

GloVe

Pennington et. al. [3]

Global statistical information of words and characters

2015

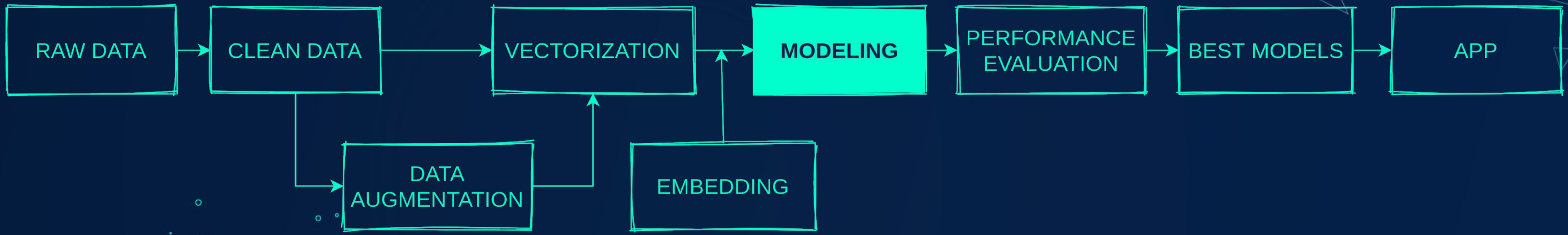
FastText

Bojanowski et. al. [4]

Character level n-gram model

Used GloVe and FastText

MODELING WORKFLOW



MODELS

MODELS

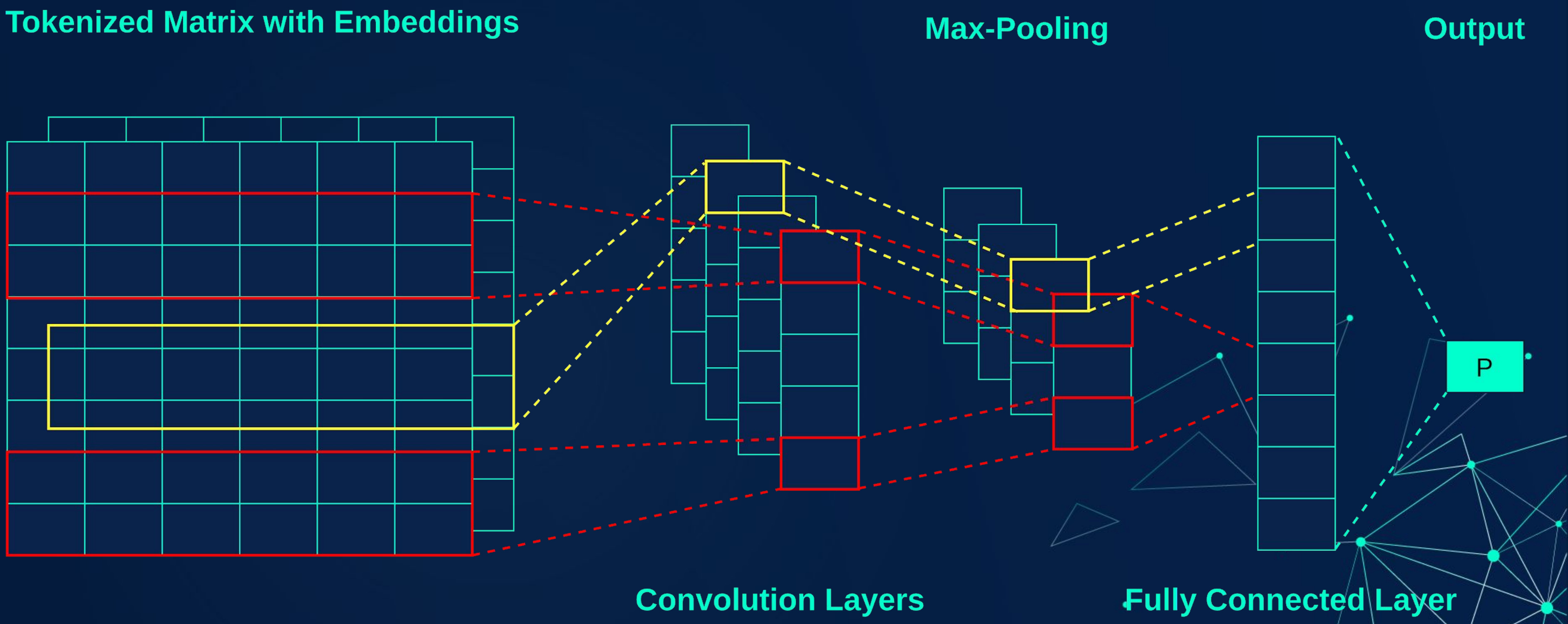
CNN

RNN

BiLSTM

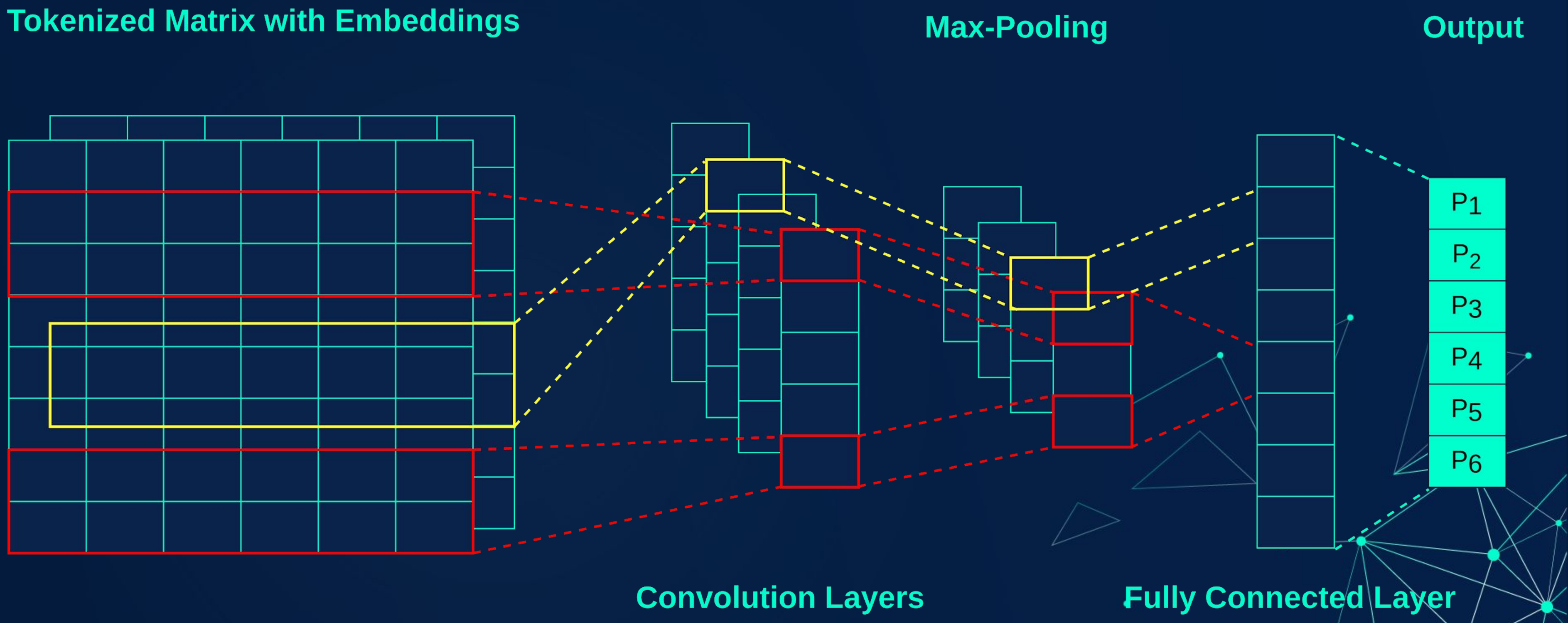
BiLSTM + Attention

MODEL ARCHITECTURE : CNN (Binary)



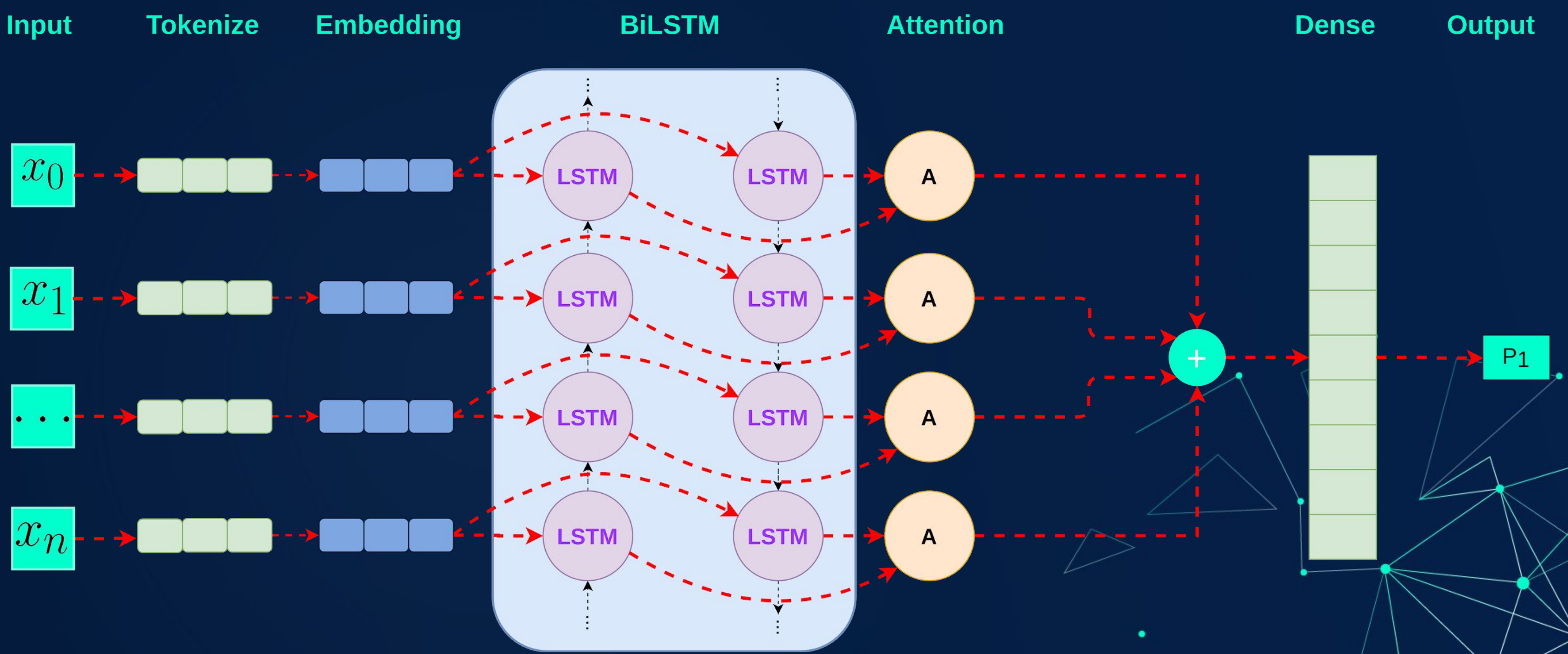
Binary Classifier with one output node

MODEL ARCHITECTURE : CNN (Multi-label)



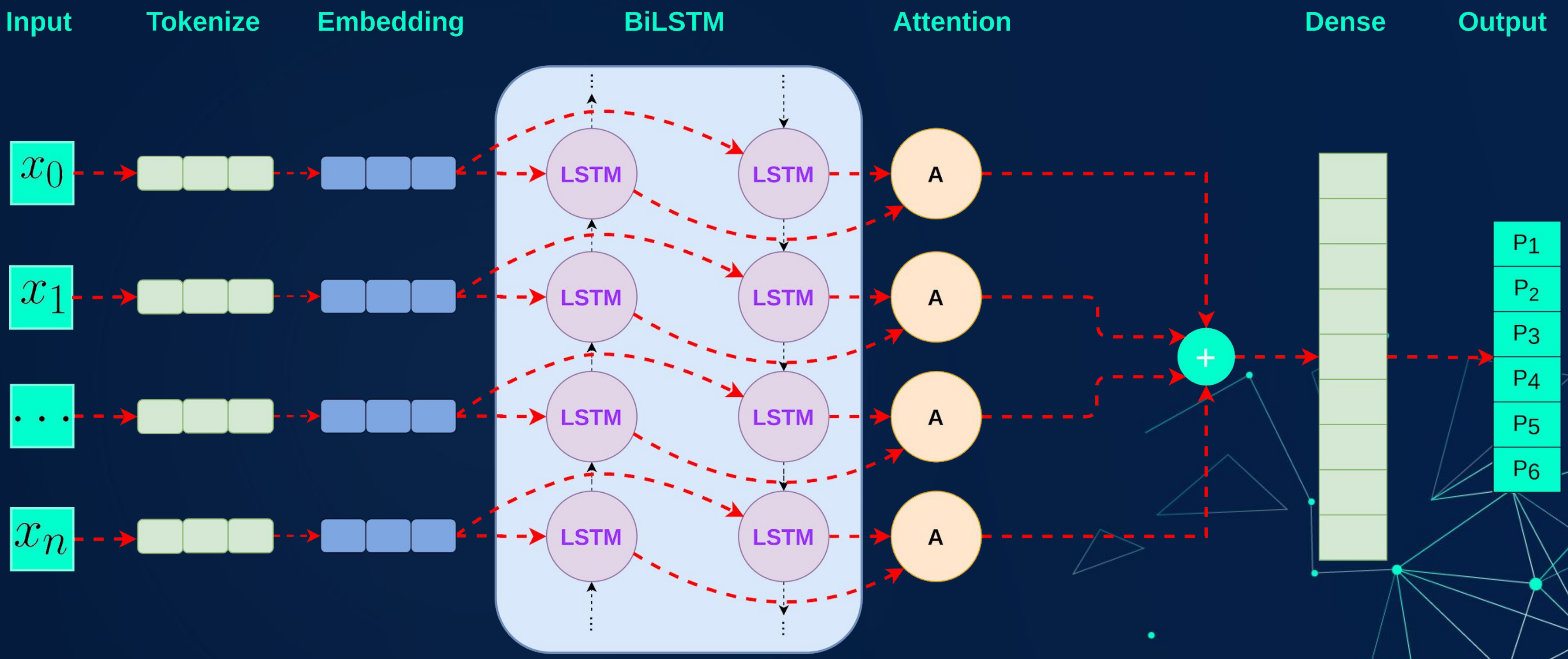
Multi-label Classifier with six output nodes

MODEL ARCHITECTURE : BiLSTM + Attention (Binary)



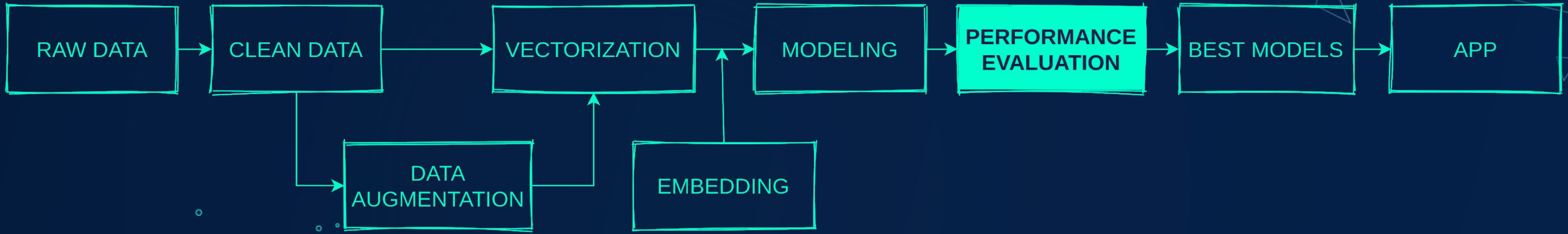
Binary Classifier with one output node

MODEL ARCHITECTURE : BiLSTM + Attention (Multi-label)



Multi-label Classifier with six output nodes

MODELING WORKFLOW



MODEL PERFORMANCE METRIC

Objectives: Minimize False Positives and **Maximize True Positive Rate**

Model performance metrics : Recall, Precision, and F1-score

$$\text{Recall (TPR)} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

MODEL PERFORMANCE METRIC

Objectives: **Minimize False Positives** and **Maximize True Positive Rate**

Model performance metrics : **Recall, Precision, and F1-score**

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

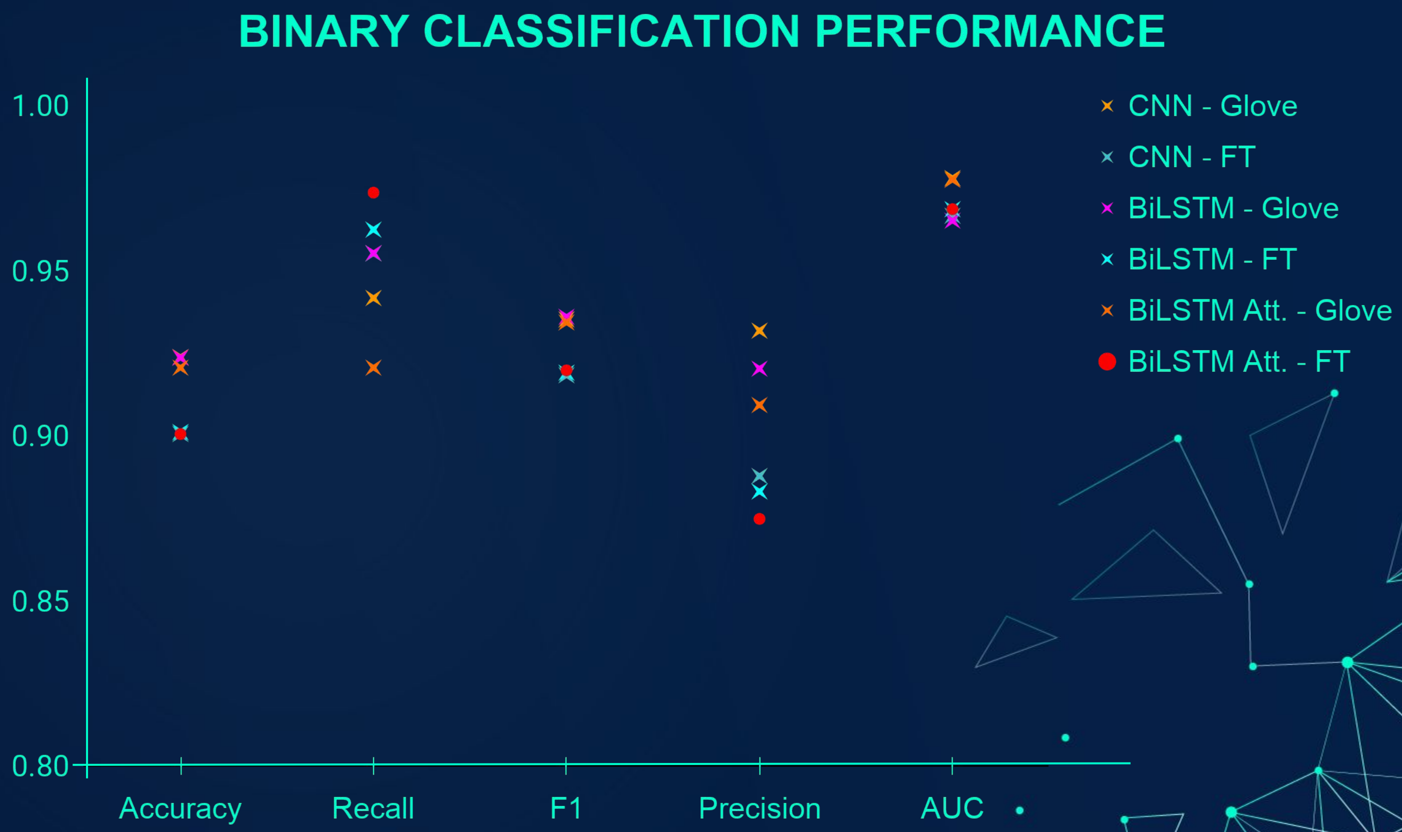
MODEL PERFORMANCE METRIC

Objectives: Minimize False Positives and **Maximize True Positive Rate**

Model performance metrics : Recall, Precision, and F1-score

$$F_1\,\text{score} = 2\,\frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

MODEL PERFORMANCE



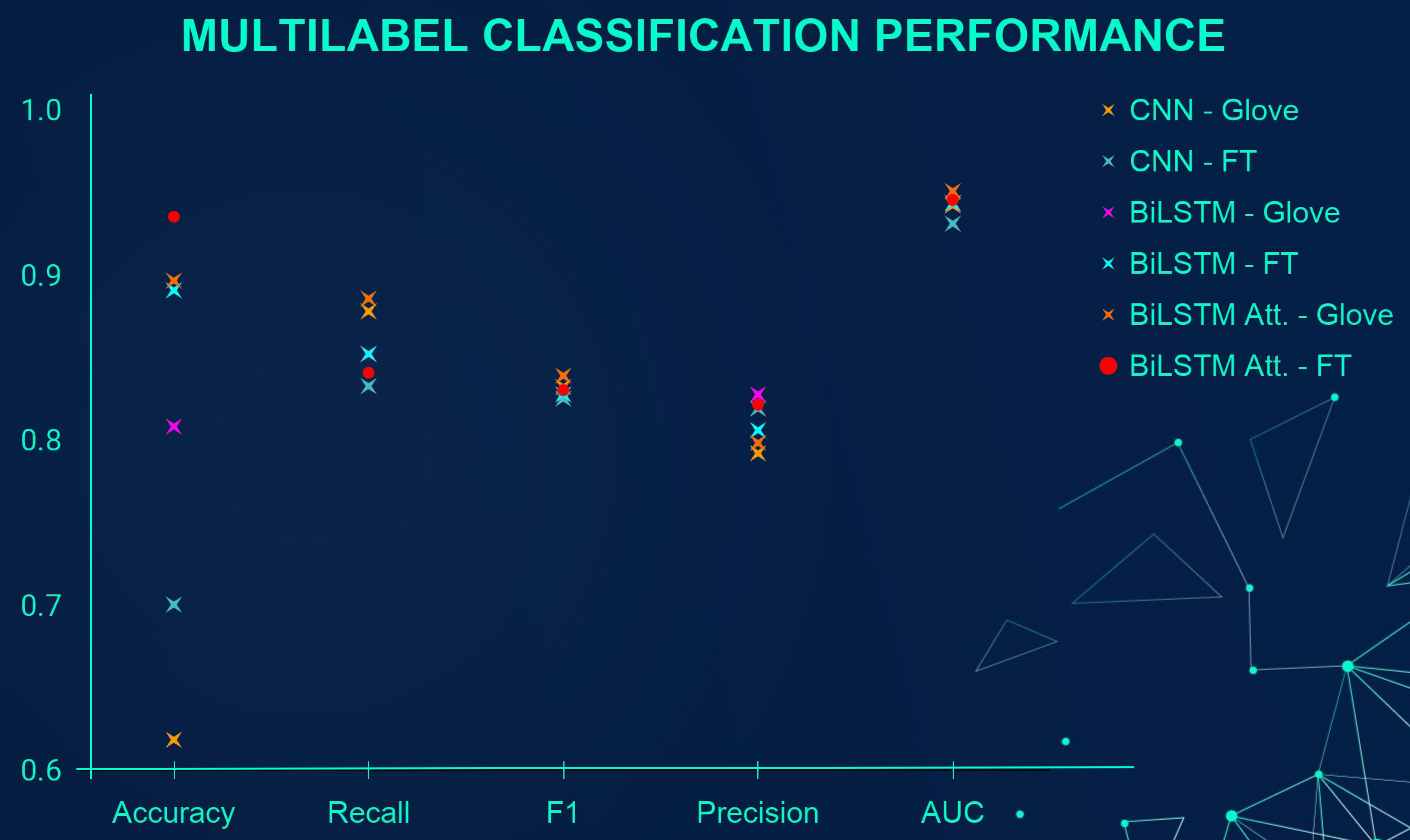
MODEL PERFORMANCE

Model Architecture	Accuracy	Recall	F1	Precision	AUC
CNN - Glove	0.9239	0.9418	0.9353	0.9319	0.9781
CNN - FT	0.9006	0.9553	0.9184	0.8879	0.9667
BiLSTM - Glove	0.9237	0.9554	0.9361	0.9204	0.9652
BiLSTM - FT	0.9012	0.9625	0.9192	0.8832	0.9687
BiLSTM Att. - Glove	0.9207	0.9207	0.9344	0.9094	0.9776
BiLSTM Att. - FT	0.9006	0.9737	0.9199	0.8749	0.9687

Worst	Best

BiLSTM + Attention with FastText embedding layer had the best Recall

MODEL PERFORMANCE



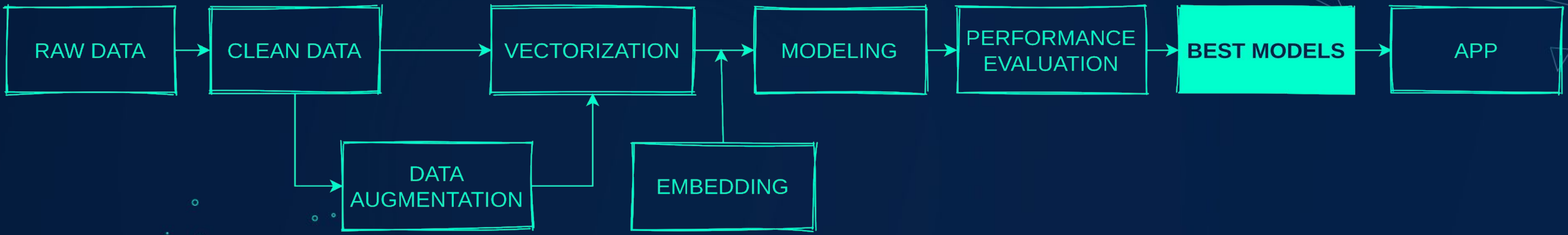
MODEL PERFORMANCE

Model Architecture	Accuracy	Recall	F1	Precision	AUC
CNN - Glove	0.6186	0.8781	0.8323	0.7921	0.9423
CNN - FT	0.7005	0.8328	0.8252	0.8192	0.9312
BiLSTM - Glove	0.8083	0.8523	0.8278	0.8278	0.9437
BiLSTM - FT	0.8908	0.8523	0.8278	0.8061	0.9437
BiLSTM Att. - Glove	0.8967	0.8858	0.8391	0.7984	0.9509
BiLSTM Att. - FT	0.9354	0.8409	0.8304	0.8216	0.9455

Worst	Best

BiLSTM + Attention with GloVe embedding layer had the best Recall

MODELING WORKFLOW



BEST MODEL SUMMARY

Binary Classifier

Model Architecture: **BiLSTM + Attention with Fast Text Embedding**
Model Performance:

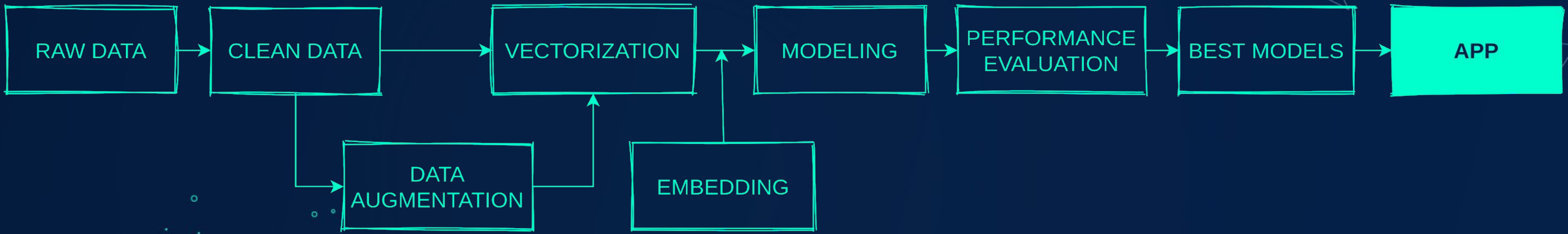
Accuracy	Recall	F1	Precision	AUC
0.9006	0.9737	0.9199	0.8749	0.9687

Multi-Label Classifier

Model Architecture: **BiLSTM + Attention with GloVe Embedding**
Model Performance:

Accuracy	Recall	F1	Precision	AUC
0.8967	0.8858	0.8391	0.7984	0.9509

MODELING WORKFLOW



DISCLAIMER

WARNING

For the upcoming demo I will be using some of the text comment from the dataset that has offensive and profane language. I would like to give you this opportunity to turn off your video if you are inclined to do so.

DEMO

Key Takeaways

Binary and multi-label classifier with Deep Neural Architecture were successful in classifying toxic text with reasonable recall

Although binary classifier has reasonably high recall and AUC score, multi-label classifier has some room for improvement

Problem of Imbalanced classes persists; Balanced data will most likely help improve the model performance

FUTURE CONSIDERATIONS

Revisit the problem with robust dataset

Investigate ways to mitigate problem of inherent bias in both, dataset and pretrained embeddings

Train models using bigger pretrained embeddings like ELMO and BERT

Fine tune DNN models with varying hidden layers to improve training performance

Deploy a REST API to serve my trained model for others to use in their application

REFERENCES

1. Civil Comment Corpus, Jigsaw-Conversation AI Data Set
2. Word2Vec Embedding
3. GloVe Embedding
4. FastText Embedding
5. GloVe: Global Vectors for Word Representation
6. Perspective API

THANK YOU

CONTACT

Email: prasoon.karma@gmail.com
Linkedin: in/karmacharya
Github: github.com/karma271
Twitter: @karma271