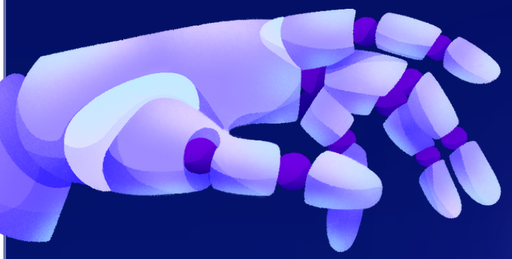
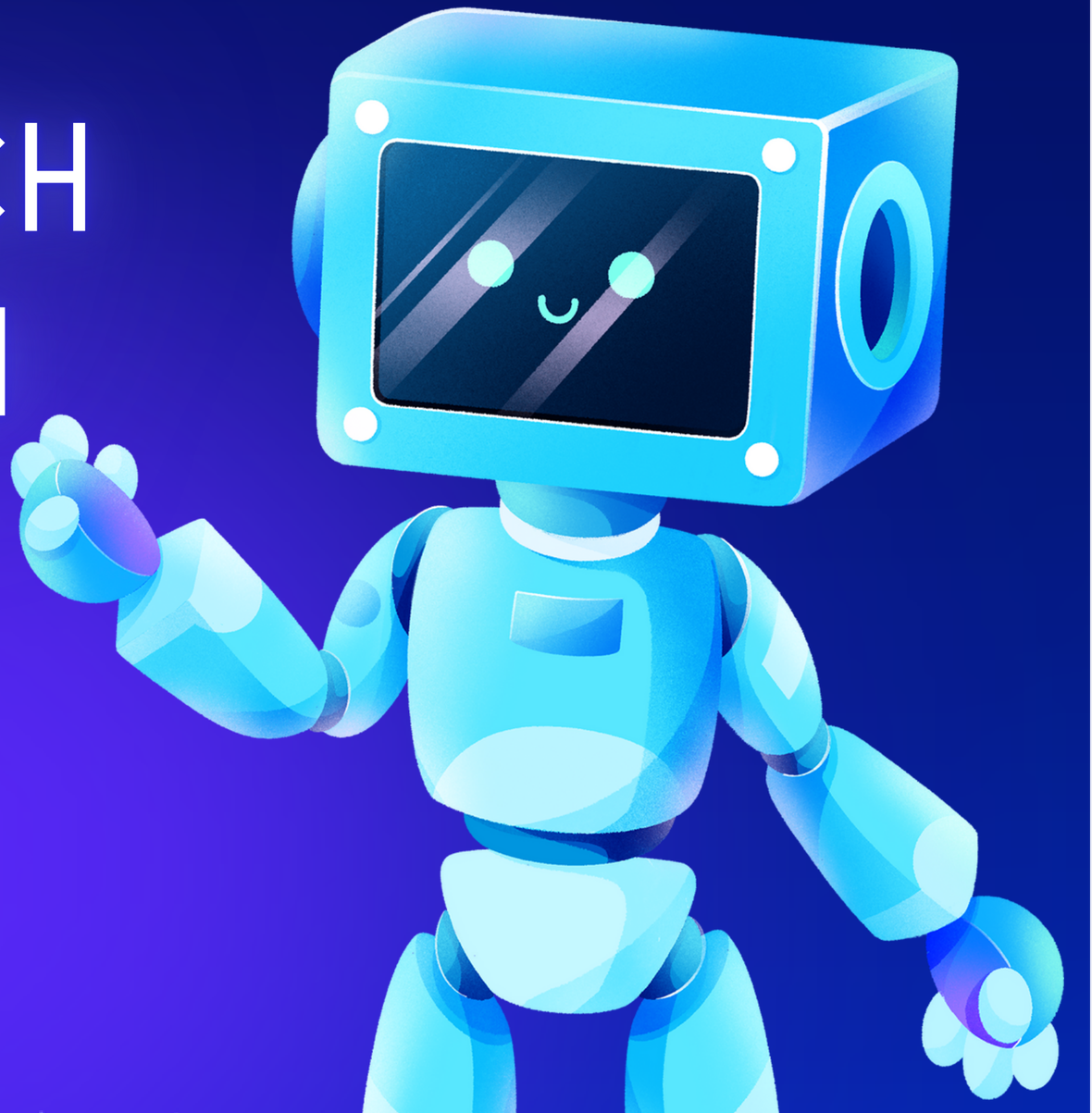


HATE SPEECH DETECTION

By
Manan Gadhiya (22070126505)
Harsh Nikam(22070126506)
Aditya Ubale(22070126512)



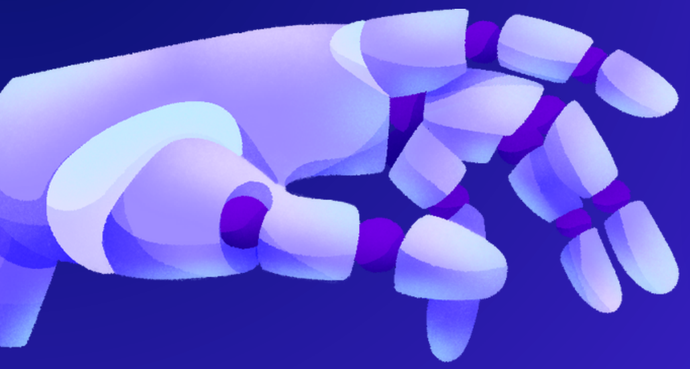
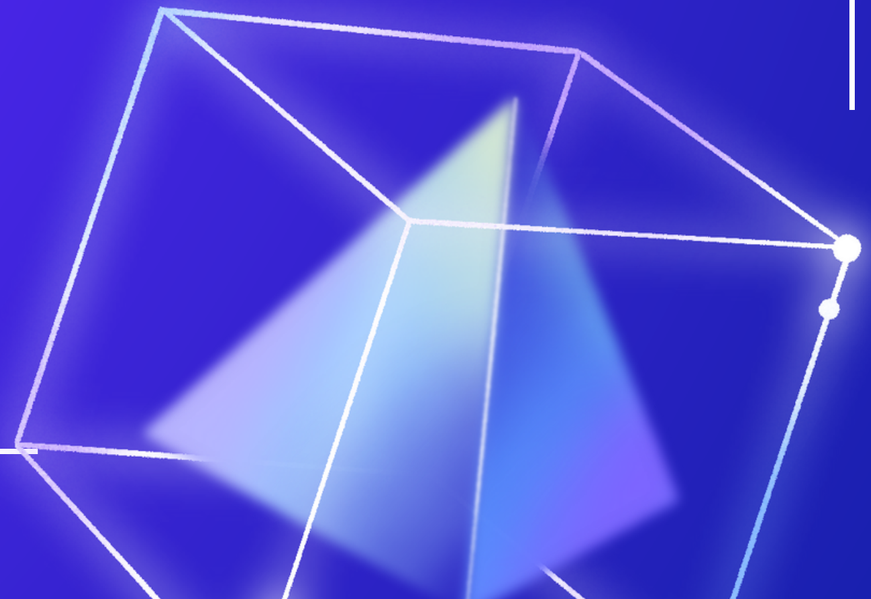


TABLE OF CONTENTS



INTRODUCTION

Hate speech detection is a critical area of natural language processing (NLP) and artificial intelligence (AI) that focuses on identifying and mitigating offensive, harmful, or discriminatory language in various forms of communication, primarily in text. In an increasingly interconnected and digital world, online platforms, social media, and digital communication channels have become breeding grounds for hate speech, making it a significant concern for individuals, communities, and society at large.

Hate speech detection systems leverage advanced NLP techniques, machine learning algorithms, and large datasets to automatically identify and classify hate speech in text data. These systems work by analyzing the linguistic and contextual features of text, including word choices, sentiment, and the context in which certain words or phrases are used.

MOTIVATION

- People refrain from expressing themselves due to toxicity on social media affecting their emotional and mental well-being.
- A system must be developed to identify such toxicity in texts.

PROBLEM STATEMENT

- To propose a model that will help users to stay away from the toxic environment that exist on the social media in the form of text.
- To propose a model for identification of toxicity i.e., toxic or non- toxic intexts.
- To propose a model with the high-rate accuracy.
- The primary problem is to design and implement robust hate speech detection models capable of accurately identifying hate speech in a variety of languages and cultural contexts.

OBJECTIVE

The main objective of this work is to develop an automated deep learning based approach for detecting hate speech and offensive language. Automated detection corresponds to automated learning such as machine learning: supervised and unsupervised learning. We use a supervised learning method to detect hate and offensive language. Classify tweets into three or four classes (like: racist, sexist, none, both) based on tweet sentiment and other features that a tweet demonstrate.

METHODOLOGY

Collection of Data

↓

Data Preprocessing

↓

Data analysis

↓

Detection of toxic word based on proposed work

↓

Performance Analysis

RESULTS

```
[23] model=DecisionTreeClassifier()
```

```
[24] model.fit(x_train , y_train)
```

```
DecisionTreeClassifier
DecisionTreeClassifier()
```

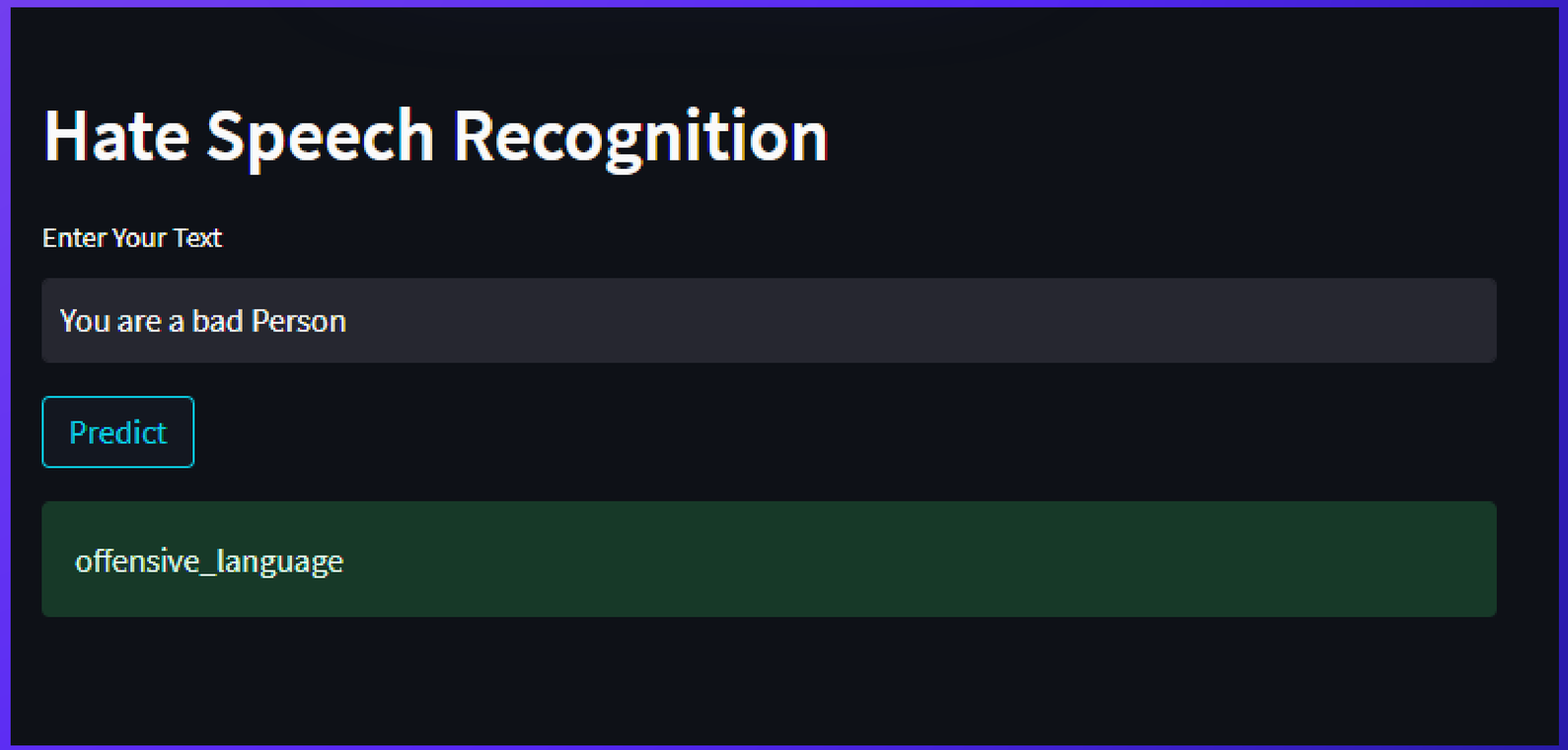
```
[25] test_data="i will save you"
df=cv.transform([test_data]).toarray()
label = model.predict(df)[0]
label_dict[label]
```

```
'no_hate_speech'
```

```
test_data="i will kill you"
df=cv.transform([test_data]).toarray()
label = model.predict(df)[0]
label_dict[label]
```

```
'hate_speech'
```

DEPLOYMENT


Hate Speech Recognition
Enter Your Text
You are a bad Person
Predict
offensive_language

LIMITATIONS

- Hate speech detection is a challenging task, and there are several limitations and complexities associated with these systems. Understanding these limitations is crucial for developing more effective solutions and setting realistic expectations.
- Hate speech detection faces a multitude of limitations, from cultural and linguistic challenges to the need to balance accuracy with privacy and freedom of expression concerns. Addressing these limitations is an ongoing effort, requiring a combination of advanced technology, ethical considerations, and regulatory compliance.

CONCLUSION

The propagation of hate speech on social media has been increasing signiticantly in recent years and it is recognised that effective counter-measures rely on automated data mining techniques. Our work made several contributions to this problem. First, we introduced a method for automatically classifying hate speech on Twitter using a deep neural network model (DCNN and MLP) that empirically improve classification accuracy. Second we did comparative analysis of our model on four publicly available datasets.

REFERENCES

1. Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). "Automated Hate Speech Detection and the Problem of Offensive Language." arXiv preprint arXiv:1703.04009.
2. Schmidt, A., Wiegand, M., & Preoţiuc-Pietro, D. (2017). "A Survey of Online Hate Speech Detection." In Proceedings of the First Workshop on Abusive Language Online.
3. Fortuna, P., & Nunes, S. (2018). "A Survey on Automatic Detection of Hate Speech in Text." ACM Computing Surveys, 51(4), 87.
4. Dinakar, K., Jones, B., Havasi, C., & Lieberman, H. (2011). "Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence." AI Magazine, 32(4), 25-29.
5. Hovy, D., Uysal, S., & Verhagen, M. (2013). "Social Media Texts for Hate Speech Detection: A Survey." arXiv preprint arXiv:1703.04009.
6. Kumar, A., Yadav, A., & Kumar, A. (2018). "Detecting Hate Speech in Multilingual Social Media." In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC-2018).
7. Facebook AI. (2020). "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes."