HATE SPEECH DETECTION USING MACHINE LEARNING

A Project report submitted in partial fulfilment of requirement for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE ENGINEERING
(*Artificial Intelligence & Data Science*)
(2021-2025)

Submitted by

KUKKAMALLA SANDESH	**21B21A4566**
A. NAVEEN KUMAR	**21B21A4576**
J ASHOK	**21B21A4565**
D SAI KRISHNA	**21B21A4569**
UMA SANJU	**21B21A4589**

Under the esteemed guidance of

MR, VISWA, M. Tech

Assistant Professor, Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

(Artificial Intelligence & Data Science)

KAKINADA INSTITUTE OF ENGINEERING & TECHNOLOGY
KORANGI-533461
(Affiliated to JNTU-Kakinada)
2021 – 2025

CERTIFICATE

This is to certify that the thesis entitled **“**HATE DETECTION USING ML**”** is being submitted by

KUKKAMALLA SANDESH

A. NAVEEN KUMAR

J ASHOK

D SAI KRISHNA

UMA SANJU

in partial requirement for the award of Degree BACHELOR OF TECHNOLOGY **in** COMPUTER SCIENCE AND ENGINEERING (**Artificial Intelligence & Data Science**) during 2021-2025 is a record of Bonafide work carried out by them under my guidance and supervision.

Project Guide	**Head of the Department**
Mr. K. Praveen, B. Tech	**Mr. S.Srinivas, M. Tech**
Assistant Professor	**Assistant Professor**
Department OF CSE	**Department of CSE**

EXTERNAL EXAMINER **INTERNAL EXAMINER**

ACKNOWLEDGEMENT

It gives us immense pleasure to acknowledge all those who helped throughout in making this project a great success.

With profound gratitude we thank **Ms. Revathi Dube, M. TECH,** Principal, *Kakinada Institute of Engineering and Technology* , forhis timely suggestions which helped us to complete this project work successfully.

Our sincere thanks and deep sense of gratitude to Mr. S. Srinivas , MTech, Head of the Department CSE, for her valuable guidance, in completion of this project successfully. We express a great pleasure to acknowledge my profound sense of gratitude to our project guide Mr. PADAGALA NAGENDRA , B.Tech, Assistant Professor in CSE Department for this valuable guidance, comments, suggestions and encouragement throughout the course of this project.

We are thankful to both Teaching and Non-Teaching staff members of CSE department for their kind cooperation and all sorts of help bringing out this project work successfully.

With Gratitude

KUKKAMALLA SANDESH	**21B21A4566**
A. NAVEEN KUMAR	**21B21A4576**
J ASHOK	**21B21A4565**
D SAI KRISHNA	**21B21A4569**
UMA SANJU	**21B21A4589**

DECLARATION

We hereby declare that the project work **“**HATE DETECTION USING ML**”** submitted to the JNTU

Kakinada, is a record of an original work done by us under the guidance of Mr. P. PAGADALA NAGENDRA BTECH, Assistant professor, Computer Science & Engineering. This project work submitted in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering. The results embodied in this project report have not been submitted to any other University or Institute for the award of any degree or diploma. This work has not been previously submitted to any other institution or University for the award of any other degree or diploma.

With Gratitude

KUKKAMALLA SANDESH	**21B21A4566**
A. NAVEEN KUMAR	**21B21A4576**
J ASHOK	**21B21A4565**
D SAI KRISHNA	**21B21A4569**
UMA SANJU	**21B21A4589**

ABSTRACT

LIST OF FIGURES

CHAPTER – 1

INTRODUCTION

1.1 INTRODUCTION

Face detection using computer vision is an essential task with impactful applications in domains such as security, digital forensics, social media, and image organization. Unlike real-time systems that require immediate feedback, this project processes static images and pre-recorded video files to identify human faces accurately.

The core functionality of the system involves automatically detecting and locating facial regions in given image or video frames using advanced deep learning techniques and computer vision algorithms. This approach enables accurate face localization across different lighting conditions, poses, and backgrounds.

Video Stream: A continuous feed of frames captured by a camera or sourced from a video file.
Face Detection Model: A pre-trained or custom-designed model capable of identifying Faces within images or video frames.
Computer Vision Libraries: Software libraries such as OpenCV, TensorFlow, or PyTorch used for image processing, model inference, and visualization.

WORK FLOW:

Frame Acquisition: Frames are continuously captured from the video stream.
Preprocessing: Frames may undergo resizing, normalization, or other preprocessing steps to prepare them for input to the Face detection model.
Face Detection: The pre-trained model processes each frame to detect Faces, usuallygenerating bounding boxes and class labels for identified Faces.
Visualization: Detected Faces are often overlaid with bounding boxes and labels on theoriginal frames for visualization.
Real-Time Display: The processed frames with Face annotations are displayed in real-time, providing immediate feedback to users or systems.

TECHNOLOGIES AND ALGORITHMS:

Deep Learning Models: Popular architectures like SSD (Single Shot Multibox Detector), YOLO (You Only Look Once), and Faster R-CNN are commonly used for real-time Face detection tasks.
Optimization Techniques: Techniques like model quantization, GPU acceleration, and model pruning are employed to optimize inference speed and efficiency.
Parallel Processing: Utilizing parallel processing capabilities of hardware (e.g., GPUs, TPUs) accelerates model inference, enabling real-time performance.

Hate Speech Detection: Concepts and Techniques

Concept 1: Definition of Hate Speech

What Constitutes Hate Speech

Hate speech refers to any form of communication that belittles, threatens, or discriminates against individuals or groups based on characteristics such as:

- **Race or ethnicity**
- **Religion**
- **Gender or gender identity**
- **Sexual orientation**
- **Disability**
- **Nationality**

Examples of Hate Speech Forms

- **Racial hate speech**: Derogatory terms targeted at a particular race.
- **Religious hate speech**: Insulting religious figures or followers.
- **Gender-based hate speech**: Misogynistic language or slurs.
- **Sexual orientation**: Homophobic or transphobic comments.

Legal vs. Social Definitions

- **Legal Definition** (varies by country): Laws typically criminalize hate speech that incites violence or hatred (e.g., Germany’s NetzDG law, US First Amendment limitations).
- **Social Definition**: Broader and subjective; includes language deemed offensive or harmful even if not illegal.

3. Concept 2: Natural Language Processing (NLP) in Hate Speech Detection

Overview of NLP

Natural Language Processing (NLP) is a branch of AI that enables machines to understand, interpret, and generate human language.

NLP Techniques in Hate Speech Detection

- **Tokenization**: Breaking text into words or phrases.
- **Stemming and Lemmatization**: Reducing words to their root forms.
- **Stop-word Removal**: Eliminating common words that carry less meaning (e.g., “the”, “is”).
- **N-gram Analysis**: Looking at sequences of n words to detect patterns.
- **Part-of-Speech Tagging**: Identifying grammatical elements to understand sentence structure.
- **Sentiment Analysis**: Identifying the tone (positive, negative, neutral) of the message.

Text Features Extracted

- Bag of Words (BoW)
- TF-IDF (Term Frequency-Inverse Document Frequency)
- Word Embeddings (Word2Vec, GloVe)

4. Concept 3: Machine Learning Models for Detection

◆ Role of Machine Learning:

Machine learning (ML) enables systems to learn patterns from text data to distinguish between hate and non-hate speech without explicit programming.

◆ Types of ML Approaches:

- **Supervised Learning**: Trained on labeled datasets to predict outputs (e.g., hate vs. non-hate).
- **Deep Learning**: Neural networks that can capture complex patterns in text (e.g., LSTM).
- **Transfer Learning**: Leveraging pre-trained models (e.g., BERT) for hate speech classification.

◆ Common Algorithms:

- **Support Vector Machines (SVM)**: Effective in high-dimensional text classification.
- **Naive Bayes**: Probabilistic model based on word frequencies; good for quick baseline models.
- **LSTM (Long Short-Term Memory)**: Captures sequence dependencies in text.
- **BERT (Bidirectional Encoder Representations from Transformers)**: Context-aware deep learning model for state-of-the-art performance.

5. Concept 4: Datasets and Challenges

Popular Datasets

- **Hate Speech 18**: Annotated Twitter dataset with hate, offensive, and neutral classes.
- **Davidson Dataset (2017)**: Contains 25K tweets labeled as hate speech, offensive language, or neither.
- **Gab Hate Corpus**: From the Gab social platform.
- **Wikipedia Detox Dataset**: Toxicity in Wikipedia comments.

Challenges in Hate Speech Detection

- **Ambiguity and Sarcasm**: Difficult for models to interpret correctly.
- **Context Dependence**: Meaning can change based on surrounding text or user.
- **Slang and Code Words**: Evasive language designed to bypass filters.
- **Imbalanced Data**: Hate speech examples are rare compared to neutral content.
- **Multilinguality**: Handling hate speech across languages.

Concept 5: Ethical Considerations and Bias in Hate Speech Detection

◆ Bias in Detection Models:

- Training data may reflect societal biases.
- Models may disproportionately flag certain groups or dialects (e.g., African American Vernacular English - AAVE).

◆ Ethical Implications:

- **False Positives**: Legitimate speech mistakenly classified as hate speech.
- **False Negatives**: Hate speech missed by the system.
- **Freedom of Expression**: Over-regulation may suppress free speech.
- **Transparency**: Black-box models (e.g., deep learning) can lack interpretability.
- **Accountability**: Responsibility for harm caused by automated moderation.

◆ Role of Human Moderation:

- AI systems should support, not replace, human moderators.
- Human review helps interpret nuanced context and ensures ethical enforcement.

---

hate speech detection using Natural Language Processing (NLP) and Machine Learning (ML):

.CODE

```
import pandas as pd
import re
import string
import pickle
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

# Load dataset
data = pd.read_csv("hate_speech_data.csv")  # Make sure to have a dataset

# Clean text
def clean_text(text):
    text = text.lower()
    text = re.sub(r'\[.*?\]', '', text)
    text = re.sub(f"[{re.escape(string.punctuation)}]", '', text)
    text = re.sub(r'\w*\d\w*', '', text)
    return text

data['clean_text'] = data['tweet'].apply(clean_text)

# Feature and target
X = data['clean_text']
```

```
y = data['label']  # 0: Not Hate, 1: Hate

# Vectorize text
tfidf = TfidfVectorizer(stop_words='english')
X_vec = tfidf.fit_transform(X)

# Train/test split
X_train, X_test, y_train, y_test = train_test_split(X_vec, y, test_size=0.2, random_state=42)

# Train model
model = LogisticRegression()
model.fit(X_train, y_train)

# Evaluate
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))

# Save model and vectorizer
pickle.dump(model, open("model.pkl", "wb"))
pickle.dump(tfidf, open("vectorizer.pkl", "wb"))
```

□ 2. app.py — Flask App to Use the Model

python

Copy

Edit

```
from flask import Flask, request, jsonify
import pickle

app = Flask(__name__)

# Load model and vectorizer
model = pickle.load(open("model.pkl", "rb"))
vectorizer = pickle.load(open("vectorizer.pkl", "rb"))

@app.route("/")
def home():
    return "Hate Speech Detection API is running!"

@app.route("/predict", methods=["POST"])
def predict():
    data = request.get_json()
    text = data.get("text")
    if not text:
        return jsonify({"error": "No text provided!"})
    
    vec_text = vectorizer.transform([text])
    prediction = model.predict(vec_text)[0]
    result = "Hate Speech" if prediction == 1 else "Not Hate Speech"
    
    return jsonify({"prediction": result})

if __name__ == "__main__":
    app.run(debug=True)
```

Concept 6: Definition of Hate Speech

Hate speech refers to any form of communication that belittles or discriminates against individuals or groups based on attributes such as race, religion, gender, ethnicity, sexual orientation, or nationality. It often includes:

- **Racial Hate Speech:** Content attacking people based on their race or ethnic background.
- **Religious Hate Speech:** Statements that degrade or incite violence against a religion.
- **Gender-based Hate Speech:** Texts that target people based on gender identity or roles.

Legal & Social Definitions

- **Legal:** Varies by country; many jurisdictions define hate speech as illegal communication that incites violence or discrimination.
- **Social:** Broader and includes speech that may not be illegal but is harmful or offensive.

Concept 7: Natural Language Processing (NLP) in Hate Speech Detection

Natural Language Processing (NLP) plays a key role in enabling computers to understand human language.

Key Techniques:

- **Tokenization:** Breaking down sentences into words or tokens.
- **Text Preprocessing:** Removing punctuation, converting to lowercase, eliminating stopwords.
- **Stemming/Lemmatization:** Reducing words to their base form.
- **Sentiment Analysis:** Determining emotional tone.
- **Vectorization:** Converting text to numerical data using techniques like TF-IDF or word embeddings.

NLP helps filter and prepare textual data for further analysis by machine learning algorithms.

Concept 8: Machine Learning Models for Detection

Machine Learning enables automated pattern recognition from data. The following types of models are widely used:

Types:

- **Supervised Learning:** Models are trained on labeled data.
- **Deep Learning:** Uses neural networks to learn high-level features automatically.
- **Transfer Learning:** Uses pre-trained models like BERT for language understanding.

Example Algorithms:

- **Naive Bayes:** Probabilistic classifier based on Bayes' theorem.
- **SVM (Support Vector Machine):** Finds the hyperplane that best separates classes.
- **LSTM (Long Short-Term Memory):** A type of neural network suited for sequential data.
- **BERT:** Pretrained transformer model that understands language context.

Concept 9: Datasets and Challenges

Popular Datasets:

- **Hate Speech 18:** A dataset for detecting hate speech on social media.
- **Twitter Hate Speech Dataset:** Annotated tweets with hate, offensive, or neutral content.

Challenges:

- **Ambiguity:** Sarcasm and context can mislead models.
- **Bias:** Models may reflect or amplify societal biases.
- **Language Diversity:** Multilingual and slang content complicates detection.
- **Context Understanding:** Difficult for simple models to grasp deeper meaning.

Concept 10: Ethical Considerations and Bias in Hate Speech Detection

Key Issues:

- **Bias in Data:** Skewed or biased training data can lead to unfair predictions.
- **False Positives/Negatives:** Incorrect classifications can result in censorship or failure to act.
- **Transparency:** Understanding how and why a model made a decision is critical.
- **Human Moderation:** Still essential to make final decisions in complex cases.

Ethics must be a central consideration in model development and deployment to ensure fairness and accountability.

Conclusion

This project showcases the integration of Natural Language Processing and Machine Learning to build a Hate Speech Detection system. It addresses critical issues like societal harm, moderation burden, and data handling. While the solution is powerful, ethical deployment and continuous model evaluation are crucial to maintaining fairness and effectiveness in real-world applications