DETECTING HATE SPEECH ON SOCIAL MEDIA USING NLP

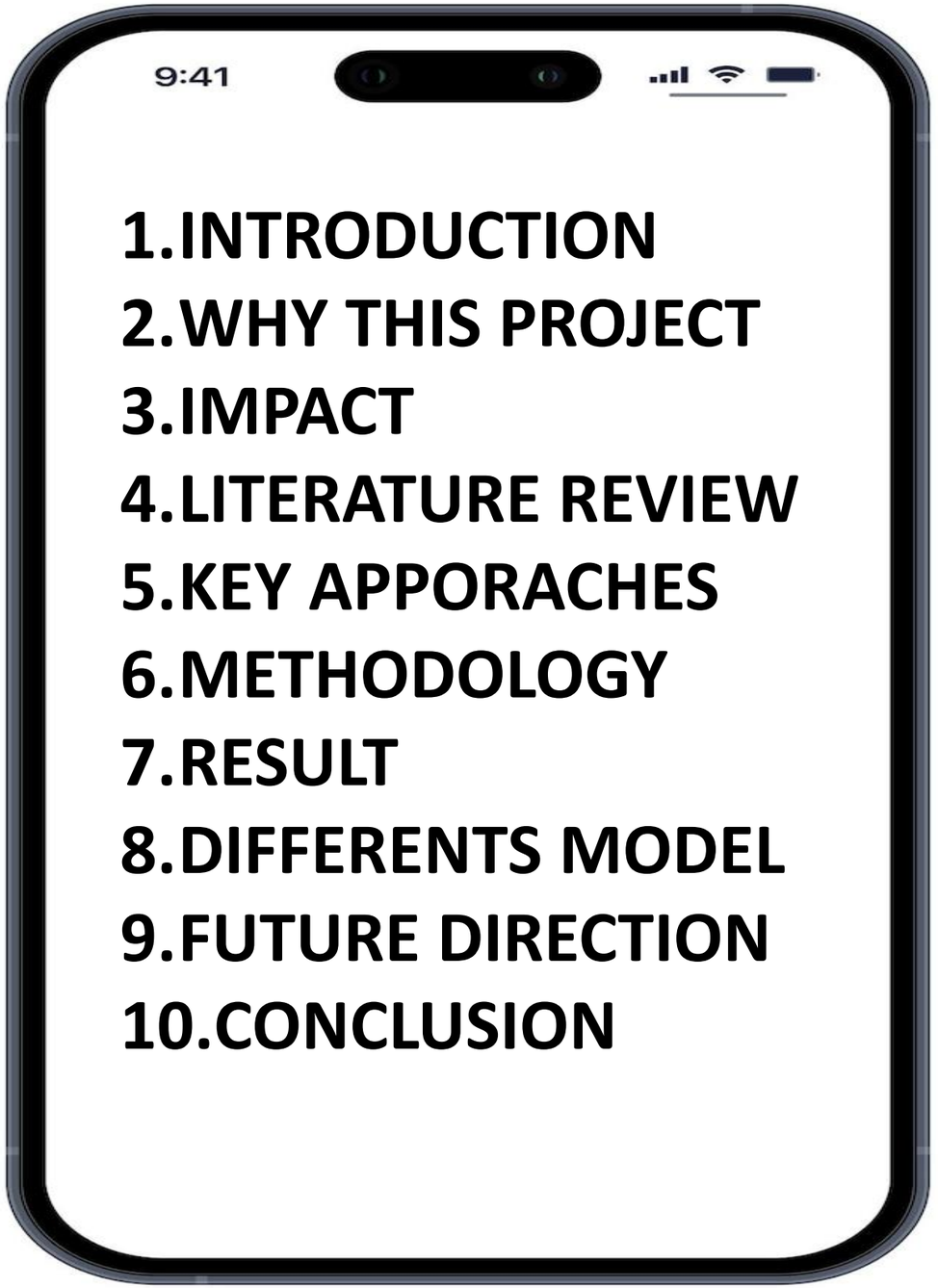
AKASH PAL,
SOUMYADEEP NASKAR,
MURLIDHAR PATRA

Under The Guidance Of
Dr. SURAJIT GIRI
Professor
Department Of Computer Science
RKMRC Narendrapur , Kolkata - 700103

AKASH PAL
ROLL NO: 6R25CMSA2001
REGN NO: A03-1142-0212-22

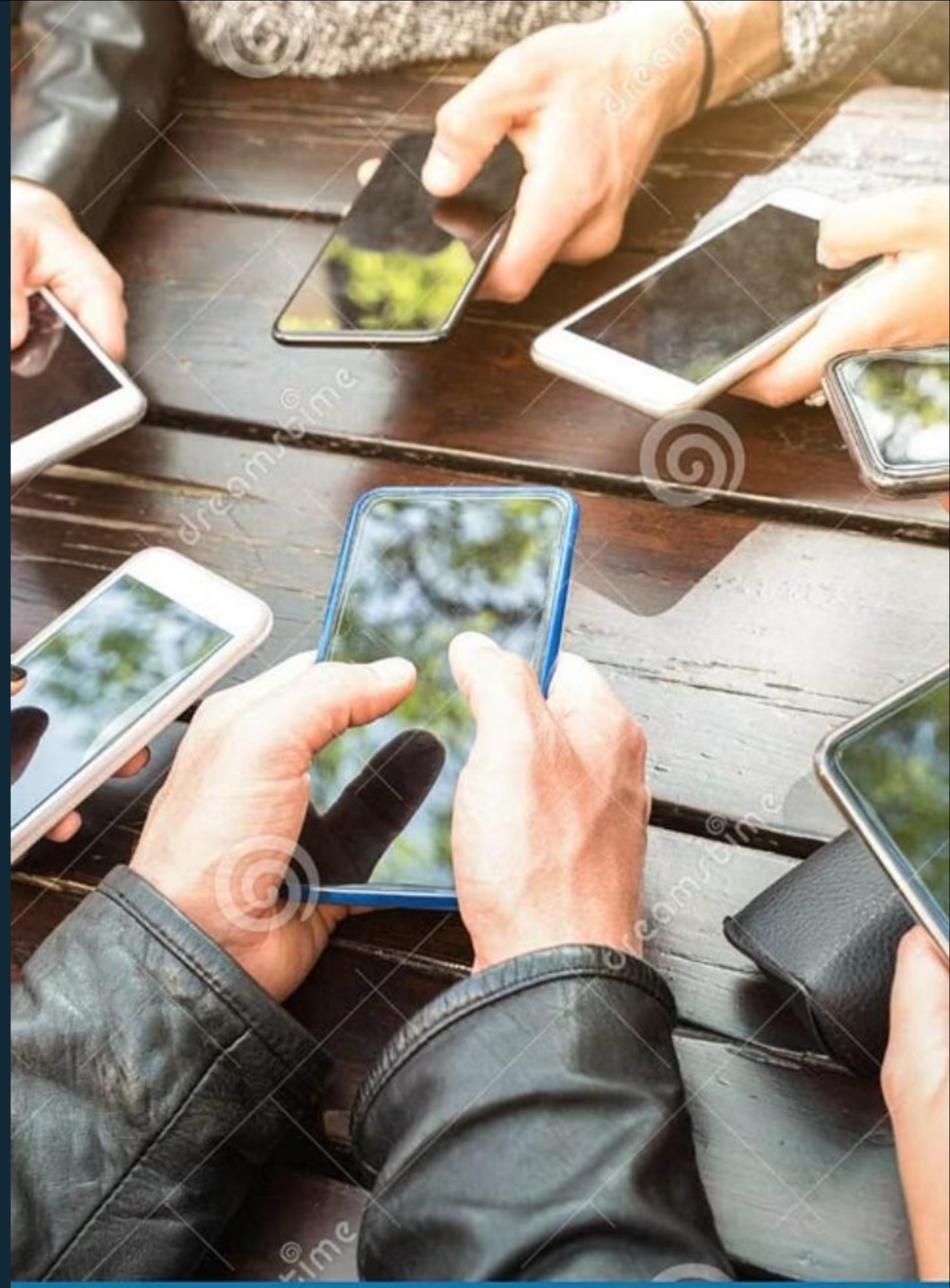
SOUMYADEEP NASKAR
ROLL NO: 6R25CMSA2011
REGN NO: A03-1122-0223-22

MURLIDHAR PATRA
ROLL NO: 6R25CMSA2015
REGN NO: A03-1132-0227-22



Detecting Hate Speech on Social Media Using NLP

This project on Natural Language Processing (NLP) to detect and mitigate hate speech on social media platforms. Online toxicity poses serious threats to vulnerable groups, spreading harmful messages that can escalate into real-world issues. Our goal is to develop accurate, effective models that can identify hate speech and help create safer digital environments. Through cutting-edge NLP techniques, we aim to enable platforms to monitor and reduce hateful content proactively.



Defining Hate Speech

Legal Definition

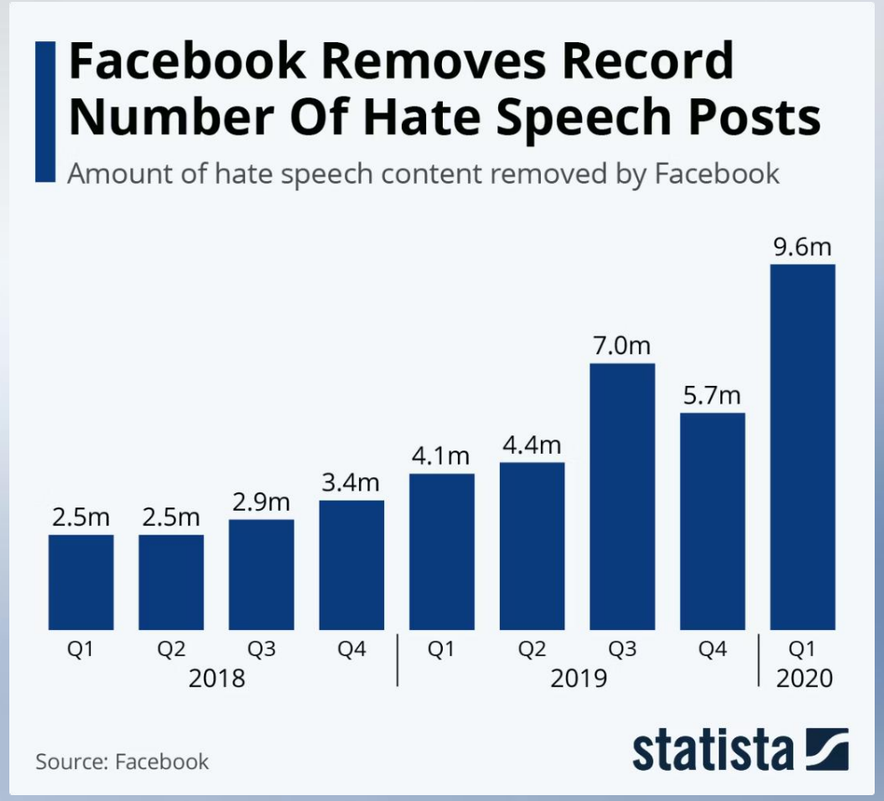
Speech that attacks individuals or groups based on protected attributes such as race, religion, nationality, sex, disability, sexual orientation, or gender identity.

Common Examples

- Use of slurs
- Stereotyping and dehumanizing language
- Calls to violence or exclusion

Free Speech vs Hate Speech

Understanding the fine line between protecting freedom of expression and preventing harmful, targeted hate speech is essential to designing responsible detection systems.



Why This Project Matters

Rising Trends

2023 saw a 38% increase in reported hate speech incidents on major social media sites, emphasizing an urgent need for intervention.

Real-World Impact

Research links spikes in online hate to a 20% rise in offline hate crimes, showing how digital behavior influences physical violence and discrimination.

Protecting Communities

Detecting hate speech early helps safeguard marginalized communities vulnerable to online harassment and abuse.

Impact on Individuals

Psychological Effects

Exposure to online hate causes anxiety, depression, and PTSD symptoms, undermining mental health.

41% of adults report experiencing online harassment, with marginalized groups disproportionately affected.

Personal Stories

- Fear of expressing identity due to targeted hate
- Increased self-censorship to avoid harassment
- Long-lasting emotional trauma from sustained abuse

Societal Implications

Democratic Discourse

Widespread hate speech erodes respectful public debate and civic engagement.

Social Trust

It undermines trust in institutions and frays social cohesion across communities.

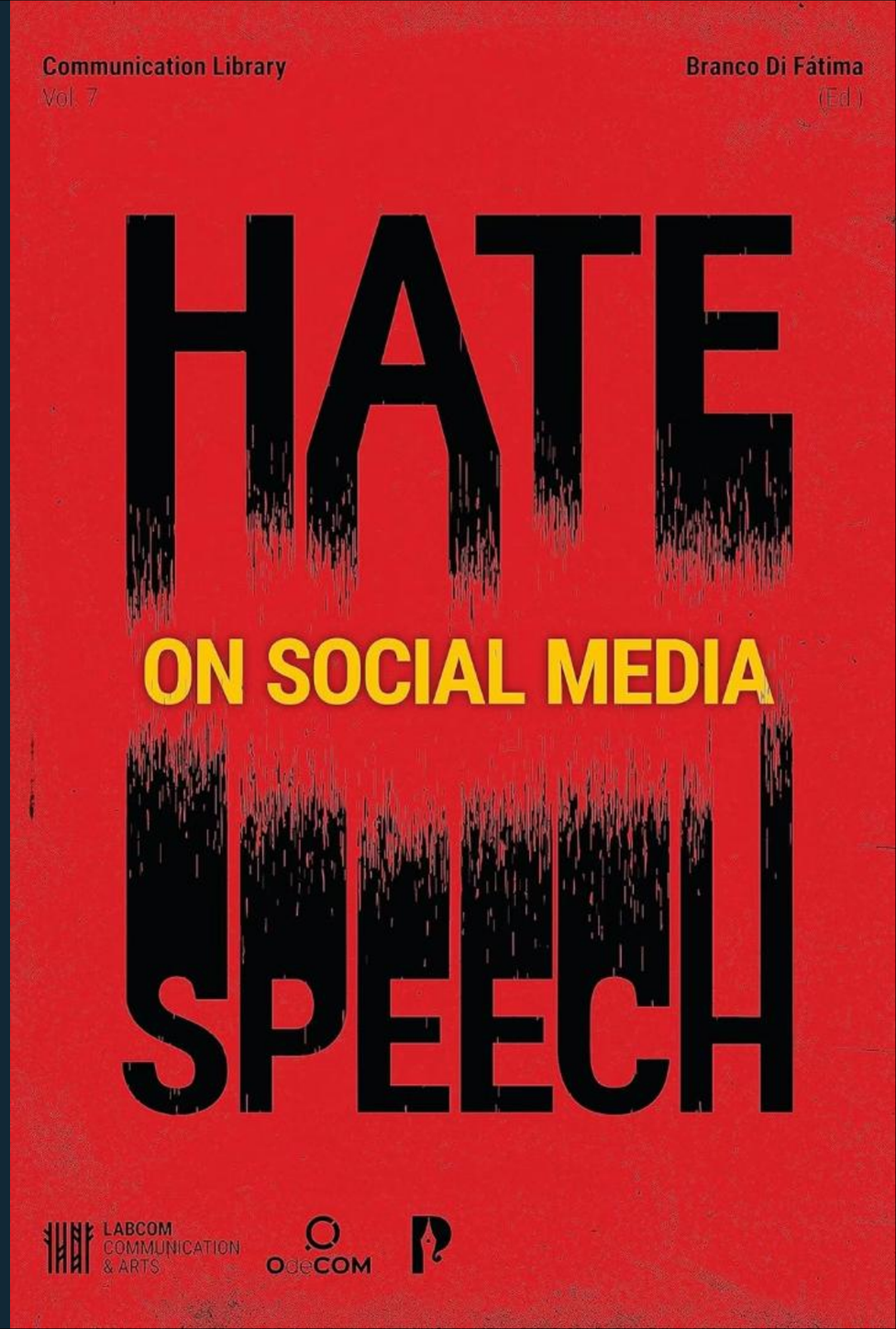
Extremism and Misinformation

Hate speech often propagates misinformation and extremist ideologies, deepening political polarization.

Literature Review: Hate Speech Detection on Social Media Using NLP

Hate speech has become a pervasive challenge on social media platforms, fostering hostility and division among users worldwide. With the rapid expansion of online communities, detecting and managing harmful content has become critical.

This presentation examines the use of Natural Language Processing (NLP) techniques developed between 2018 and 2024 for automated hate speech detection and mitigation. We explore key methodologies, datasets, and emerging trends from recent academic research, aiming to understand how technology can support safer online environments.



Key Approaches and Datasets for Hate Speech Detection

Lexicon-based Methods

Early hate speech detection relied on predefined lexicons of offensive terms, automating detection via keyword matching. However, these methods struggled with subtlety and evolving language.

DOI: 10.1145/3219819.3220084

Machine Learning Techniques

Supervised classifiers such as SVM and Logistic Regression use annotated datasets with features like TF-IDF and sentiment scores to identify hate speech effectively.

DOI: 10.18653/v1/W18-5102

Deep Learning Models

Advancements incorporate CNNs, RNNs, and Transformer models, including BERT and RoBERTa, fine-tuned for classification with datasets like Hateval and Stormfront.

Challenges remain around data imbalance and capturing context.

DOI: 10.18653/v1/N19-1256

Emerging Trends and Future Directions

Context-Aware Models

Innovative methods integrate user profiles and social networks using Graph Neural Networks, improving relational context understanding.

DOI: 10.1145/3394486.3403397

Multilingual Detection

Efforts to detect hate speech in low-resource languages leverage cross-lingual transfer learning to broaden inclusivity internationally.

DOI: 10.18653/v1/2020.acl-main.474

Explainable AI & Mitigation

Explainable AI promotes transparency by highlighting model reasoning and addressing biases. Simultaneously, ethical mitigation involves content moderation and user education strategies.

DOI: 10.1145/3461123.3461273

Project Methodology

1

Data Collection

Gathering labeled social media posts using established datasets like Hate Speech Detection Dataset on Kaggle.

2

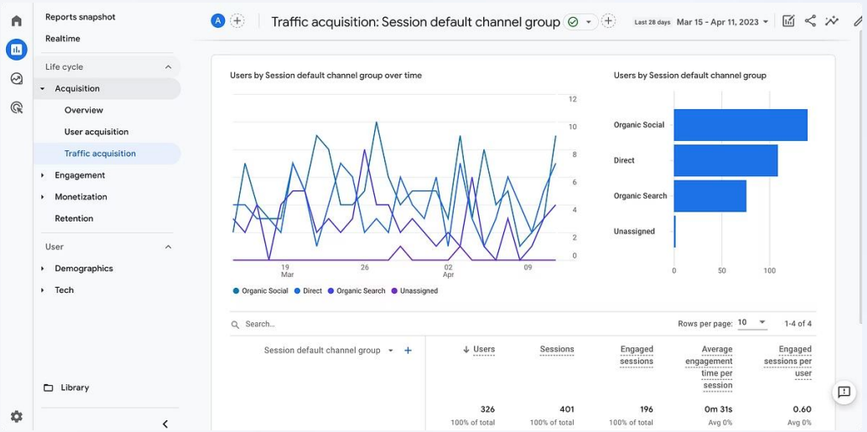
Preprocessing

Cleaning text data with tokenization, stemming, stop word removal, and vectorization for model input.

3

Model Training

Using supervised learning techniques to build detection models.



Understanding Hate Speech Detection: An Analysis of a Twitter Dataset

This presentation explores a unique dataset consisting of 24,783 tweets, each labeled as hate speech, offensive language, or neither. The dataset provides valuable insight into online discourse, highlighting how hate and offensive content differ from neutral interactions.

Our goal is to understand the dataset's structure, the distribution of labels, and its potential applications in training hate speech detection models for improved social media moderation and research.

Column Name	Description
Count ---------→	Number of users who labeled the tweet
hate_speech_count --------→	Number of annotators who labeled it as hate speech
offensive_language_count ->	Number of annotators who labeled it as offensive language
neither_count ----------→	Number of annotators who labeled it as neither
Class ----------→	Final label: • 0 = Hate Speech • 1 = Offensive Language • 2 = Neither
Tweet ----------→	The actual tweet content

Dataset Composition and Characteristics

Total Tweets

The dataset contains 24,783 tweets, each uniquely authored and distinct in tone, punctuation, and style.

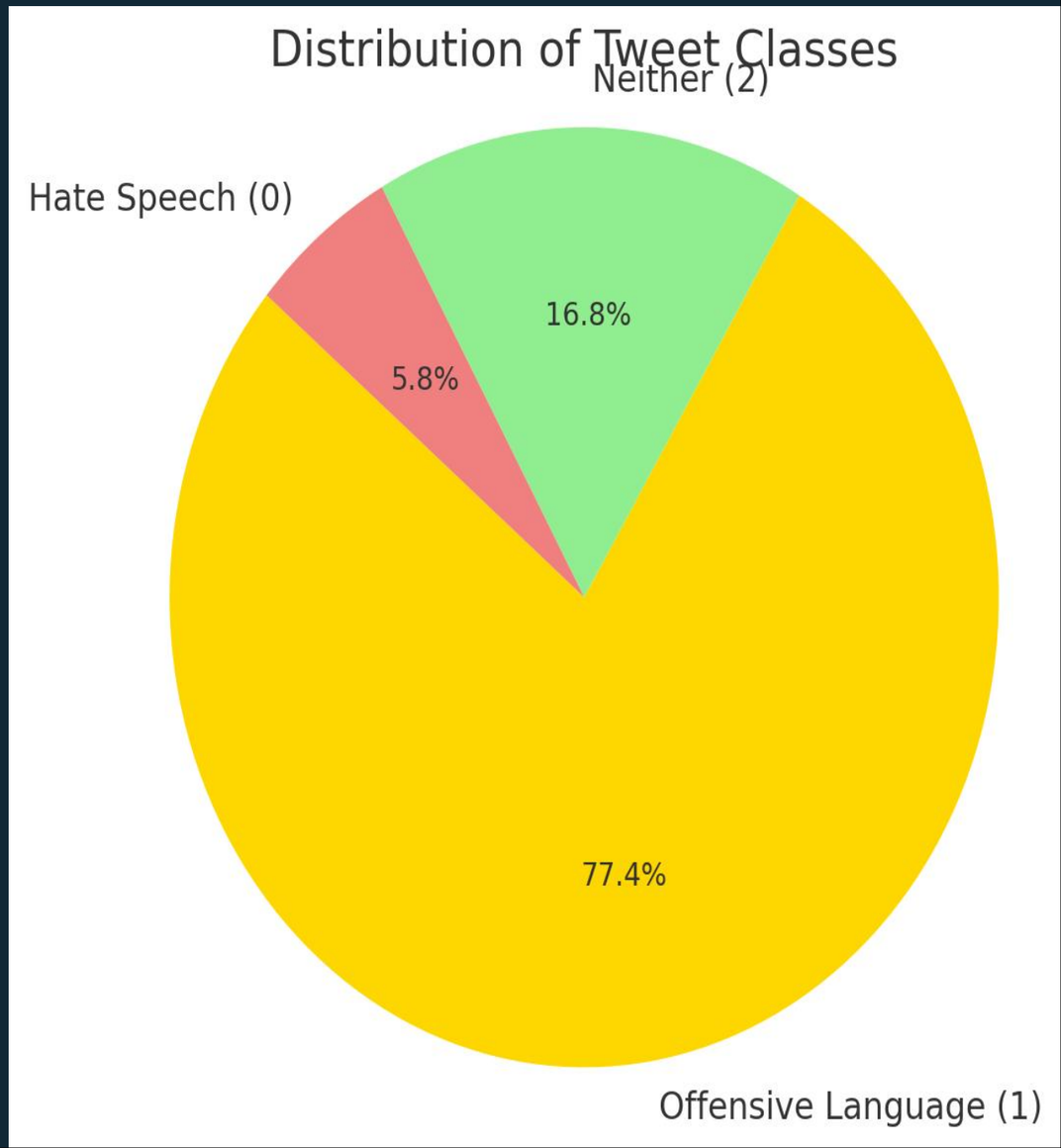
Label Categories

- Hate Speech (Class 0)
- Offensive Language (Class 1)
- Neither (Class 2)

Offensive language dominates the dataset, while hate speech comprises the smallest portion.

Annotation Details

Each tweet includes multiple annotation counts that contribute to a final class label, providing layered insights into content classification.



Text Preprocessing & Modeling: From Raw Data set

Transforming raw text data into meaningful insights requires a comprehensive approach. This presentation explores essential preprocessing techniques, word embedding methods, deep learning models, and future advancements in natural language processing (NLP). Understanding these key concepts is fundamental for building accurate, efficient, and interpretable NLP systems.

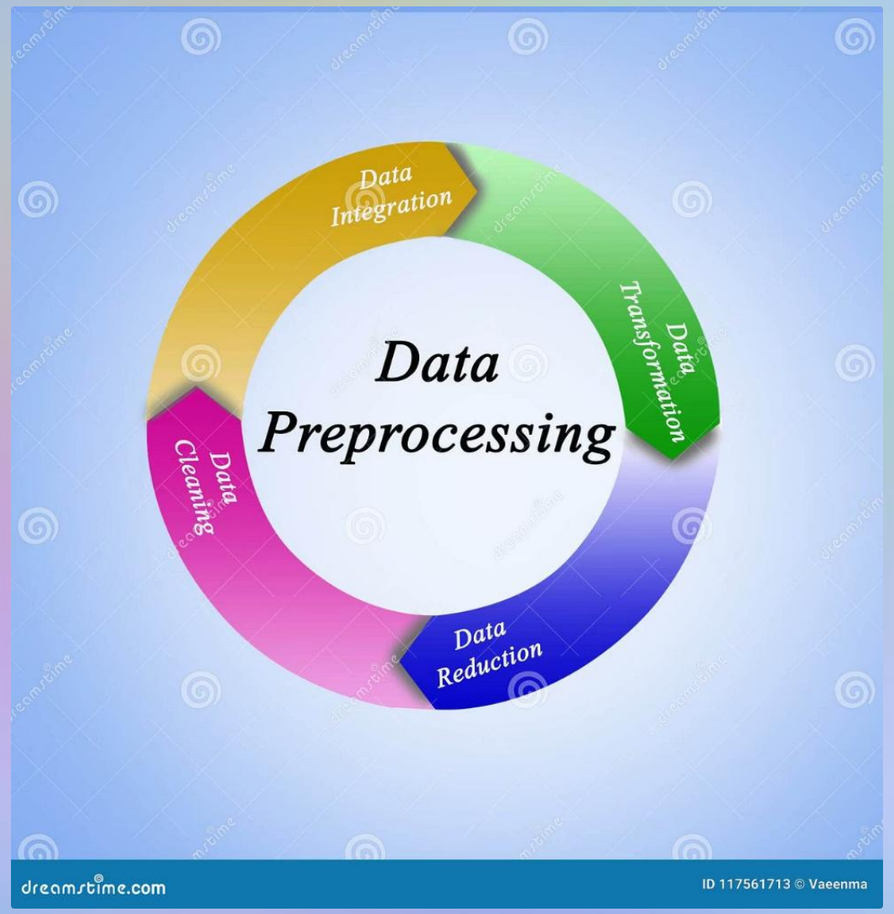
Introduction to Text Preprocessing

Why Preprocess Text?

Text preprocessing reduces noise in data to improve the accuracy and efficiency of models. It standardizes input for better analysis.

Essential NLP Step

Preprocessing forms the backbone of NLP pipelines, laying a solid foundation for downstream tasks like sentiment analysis and classification.



Preprocessing Steps: Cleaning and Normalization

1

3.1 Lowercasing

Converts all text to lowercase to ensure uniformity and treat words like "The" and "the" as identical.

2

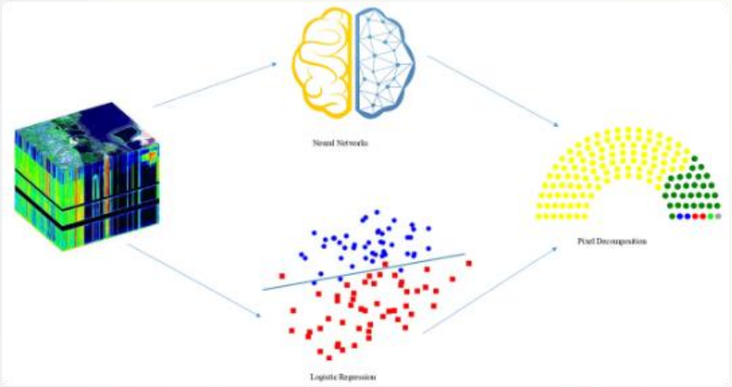
3.2 Tokenization

Splits text into individual tokens or words, enabling detailed linguistic analysis.

3

3.3 Non-Alphabetic Removal

Filters out punctuation , emojis , numbers, and special characters, focusing on meaningful content.



Preprocessing Steps: Reducing Dimensionality

3.4 Lemmatization

Transforms words to their base form (working -> work) , improving model generalization by grouping variants of the same word.

3.5 Stop Word Removal

Eliminates common filler words to sharpen focus on informative terms. Like as, is, am are etc.

3.6 Filtering

Removes rarely used or overly frequent words to reduce noise and improve overall model performance.

3.7 Joining / Chunking

Joining refers to the process of combining separate tokens (words) back into a coherent sentence after tokenization .

Example :

[‘I’ , ‘Love’ , ‘NLP’ }

Joining will reconstruct it as:

“I Love NLP”

Chunking, also known as shallow parsing, involves grouping individual words into meaningful phrases based on their part-of-speech (POS) tags. The most common chunks include noun phrases (NP), verb phrases (VP), and prepositional phrases (PP). Unlike full parsing, chunking does not analyze the entire grammatical structure, but focuses on phrase-level groupings.

Example:

For the sentence:

"The quick brown fox jumps over the lazy dog"

Chunking may produce:

- **[NP The quick brown fox]**
- **[VP jumps]**
- **[PP over]**
- **[NP the lazy dog]**

This helps in extracting subjects, actions, and objects from text.

Word Embeddings: Representing Text as Vectors

Word Embedding is a technique in Natural Language Processing (NLP) where words are represented as dense vectors (real-numbered arrays) in a continuous vector space. Unlike one-hot encoding, word embeddings capture the semantic meaning and relationships between words.

Why Use Word Embeddings?

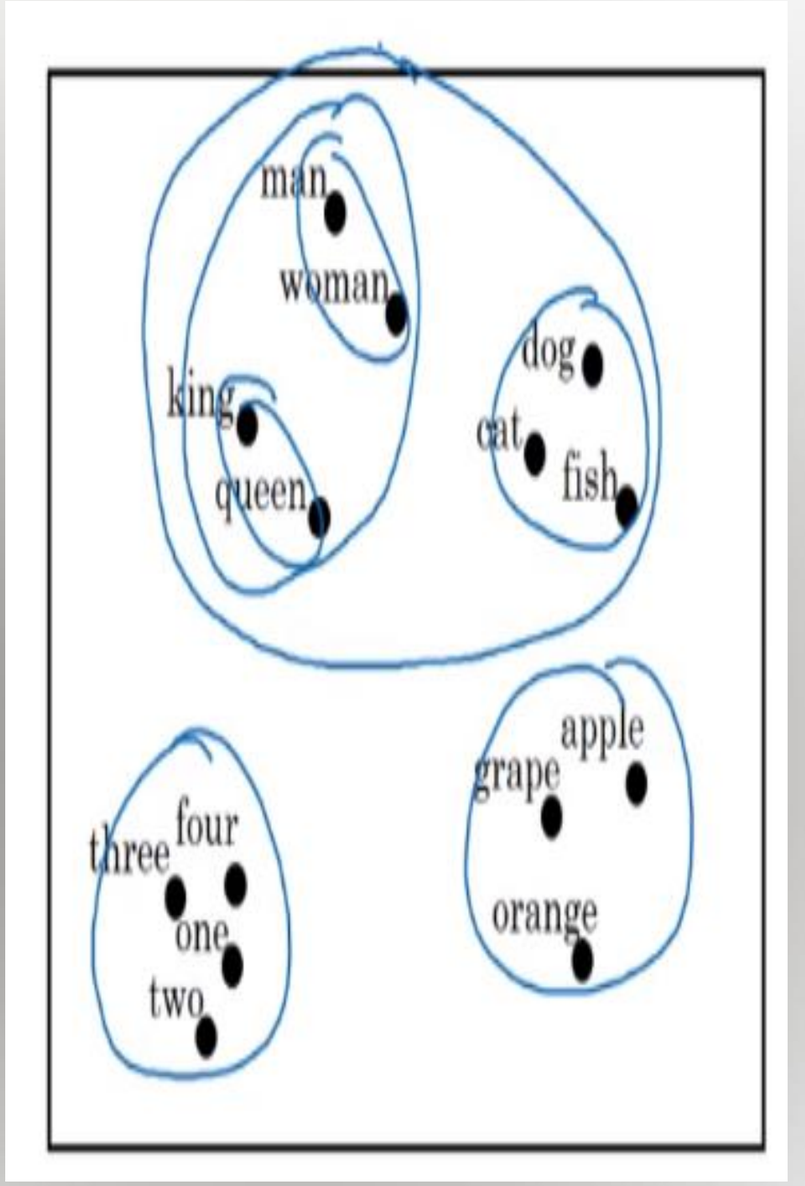
- **Reduces dimensionality** compared to one-hot encoding.
- Improves model performance in tasks like sentiment analysis, machine translation, etc.
- Captures **contextual similarity** (e.g., king – man + woman ≈ queen)

Example:

Suppose we have the words:

- **king** → [0.25, 0.78, -0.33, ...]
- **queen** → [0.24, 0.76, -0.31, ...]
- **apple** → [0.85, -0.12, 0.44, ...]

Here, the vectors for "king" and "queen" are more similar to each other than to "apple", indicating **semantic closeness**.



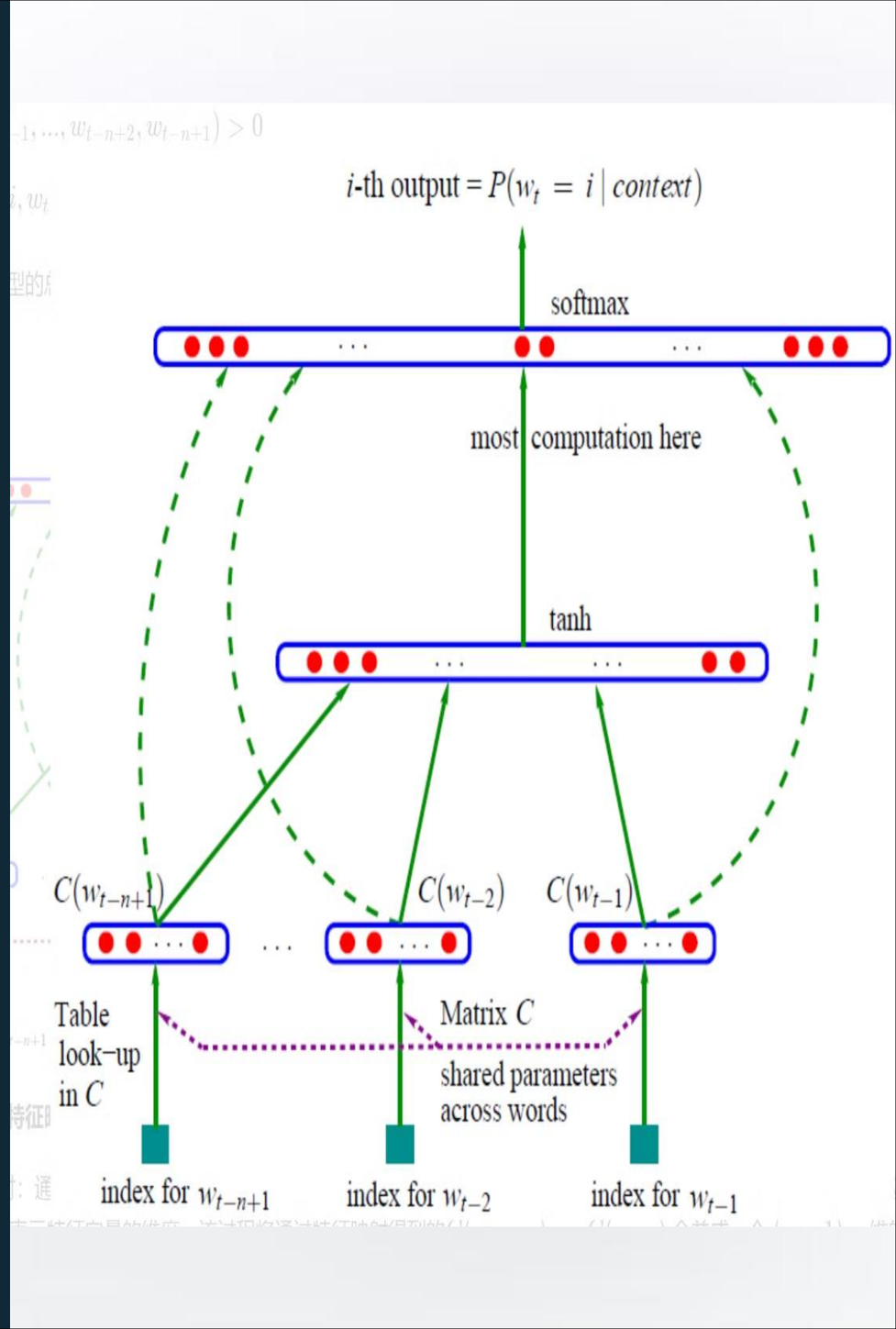
Word Embedding Techniques: The Classics

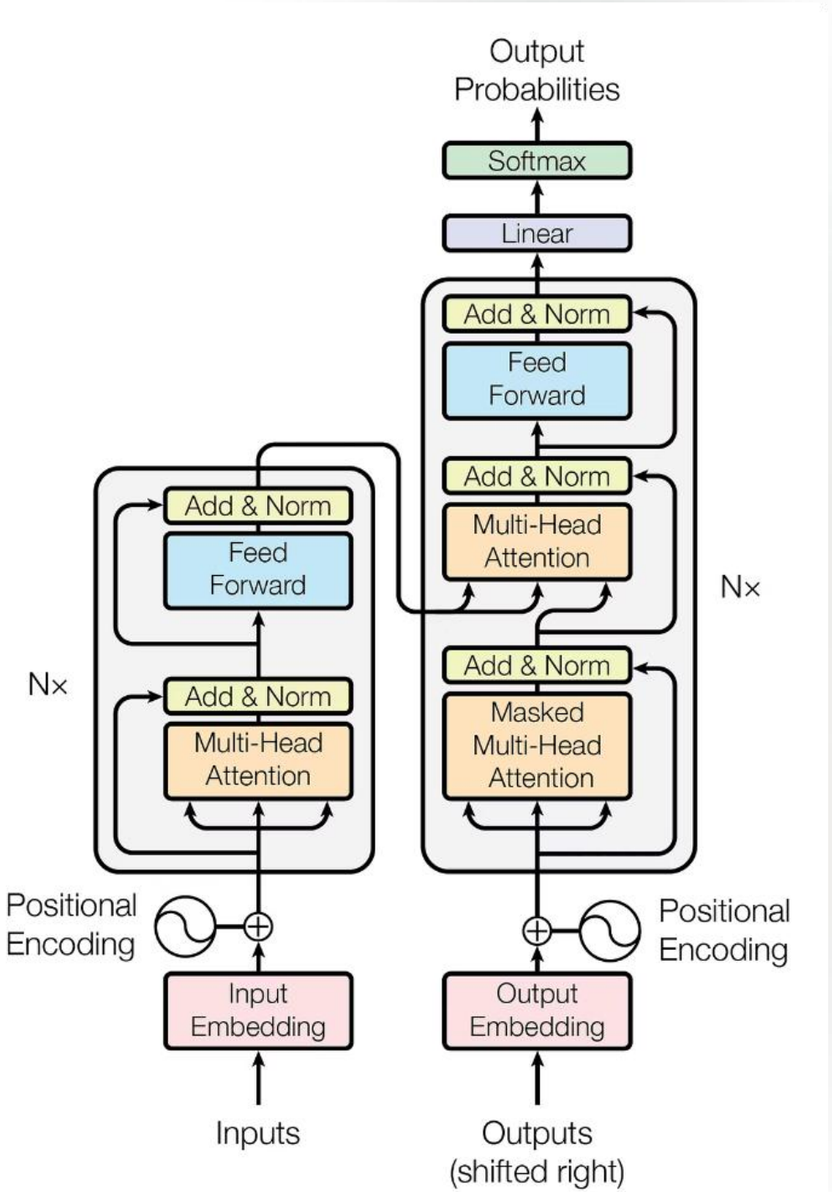
Word2Vec

Uses CBOW and Skip-gram architectures to predict words in a context window, trained on massive datasets.

GloVe

Combines global word co-occurrence statistics via matrix factorization for efficient embedding generation.





Modern Word Embeddings: Context is Key

BERT revolutionizes word embeddings by understanding words in context using bidirectional transformers. Unlike static embeddings, BERT's dynamic representations capture meaning based on surrounding text, achieving state-of-the-art results in many NLP benchmarks. This approach enables nuanced interpretation of polysemous words and complex sentence structures.

Model Application: CNN for Text Classification

Feature Extraction

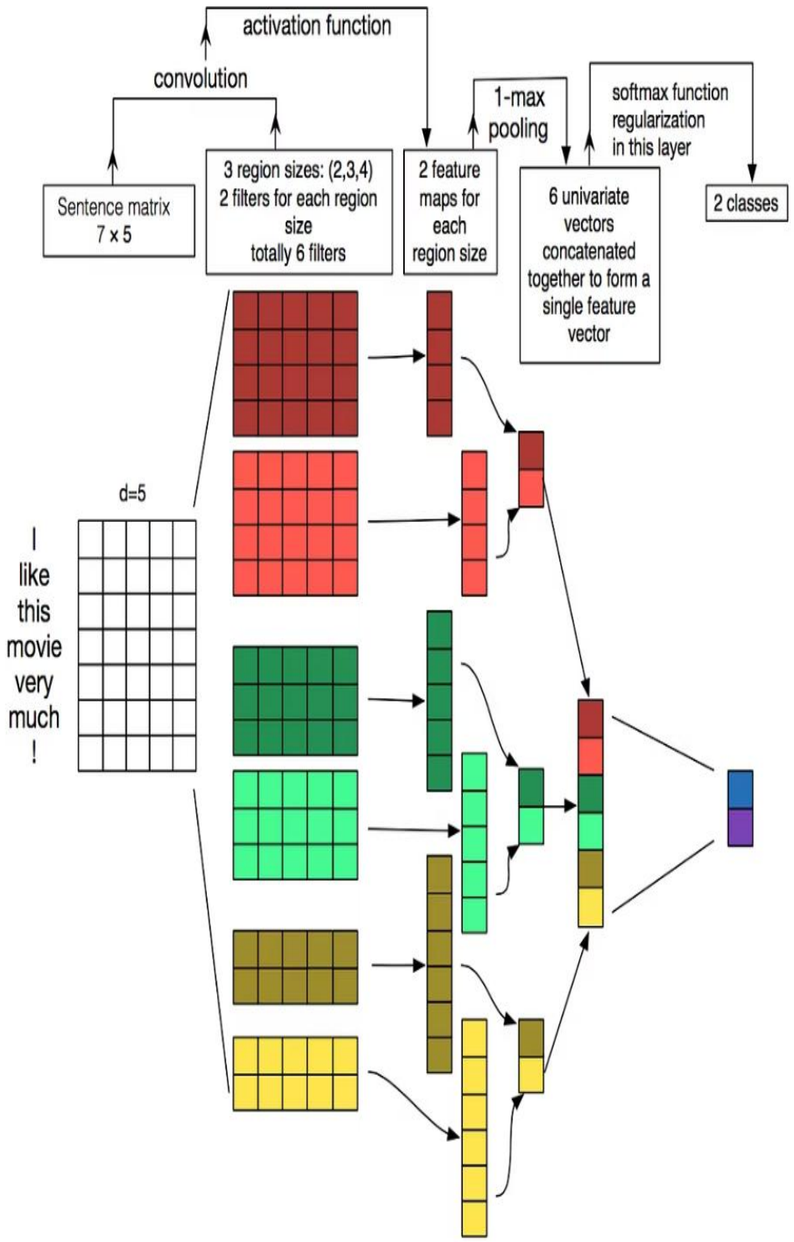
CNNs extract local patterns and n-gram features from text sequences effectively.

Sentiment & Topic Identification

Excels in classification tasks such as spam detection and sentiment analysis.

Real-World Use

Used to identify spam emails, improving filtering accuracy and user experience.



NLP Models

Baseline Models

- Logistic Regression
- Naive Bayes
- 75-80% accuracy on initial tests

Advanced Models

- BERT, RoBERTa, Transformer architectures
- Achieve 90%+ accuracy on benchmarks
- Bias mitigation techniques applied

Evaluation metrics such as precision, recall, F1-score, and AUC-ROC quantify model effectiveness and fairness.

Model Evaluation and Results

Performance Metrics

Comparing accuracy, precision, recall, and F1-scores across various models to identify the best performers.

Error Analysis

Identifying common misclassifications to improve model robustness and reduce false positives and negatives.

Visualizations

Using confusion matrices and validation loss vs training and validation accuracy vs training graph

Class	Precision	Recall	F1-Score
0	0.37	0.24	0.29
	0.91	0.92	0.92
2	0.75	0.80	0.78

Training vs Validation Graph in Machine Learning (NLP Context)

Definition:

A **Training vs Validation graph** shows how a model learns over time by plotting performance metrics (like **accuracy** or **loss**) on both the **training** and **validation** datasets across **epochs**.

Common Metrics Plotted:

- **Training Accuracy / Loss**: Measures performance on the data the model is learning from.
- **Validation Accuracy / Loss**: Measures generalization to unseen data (important for real-world performance).

Purpose:

- To monitor **model learning progress**
- To detect issues like **overfitting** or **underfitting**

Example Interpretation:

Ideal Case:

Both training and validation accuracy **increase steadily** and stabilize with **minimal gap**.

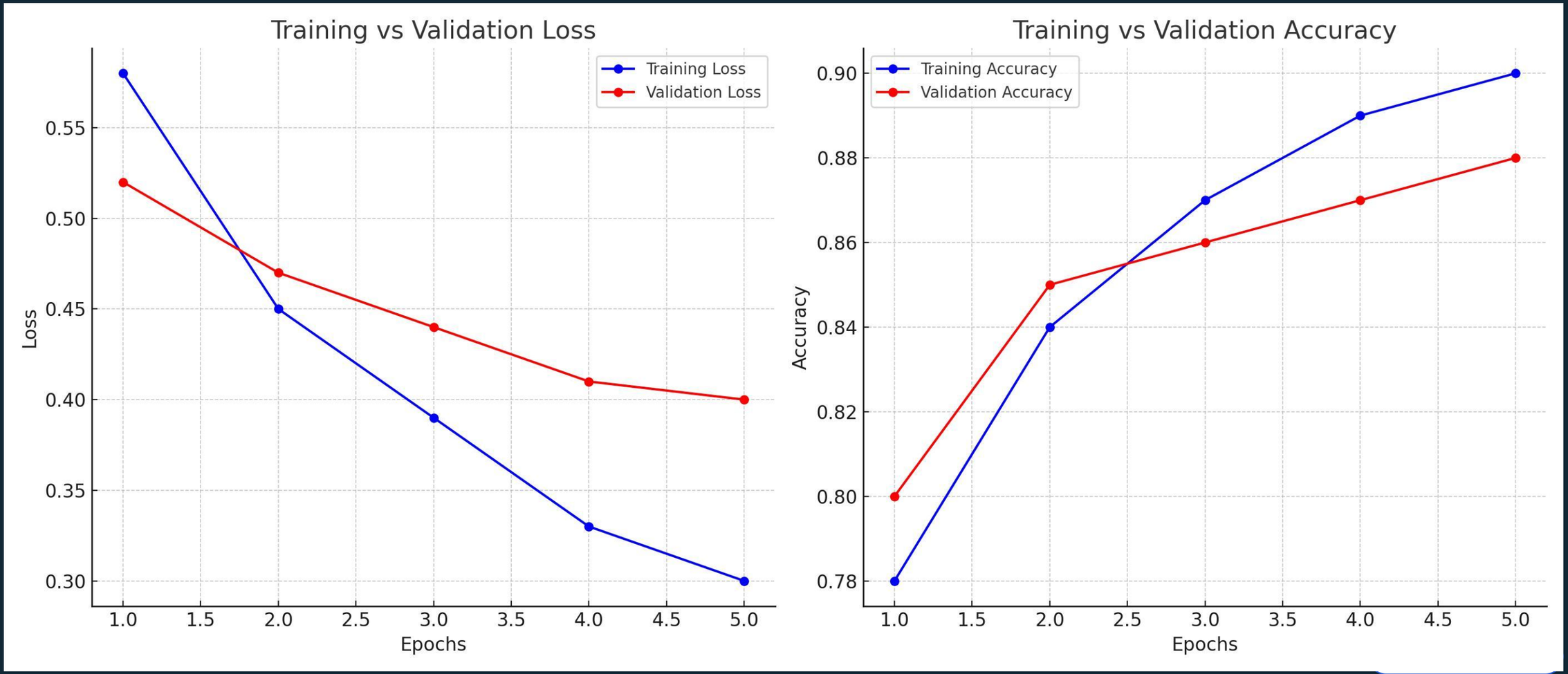
Overfitting:

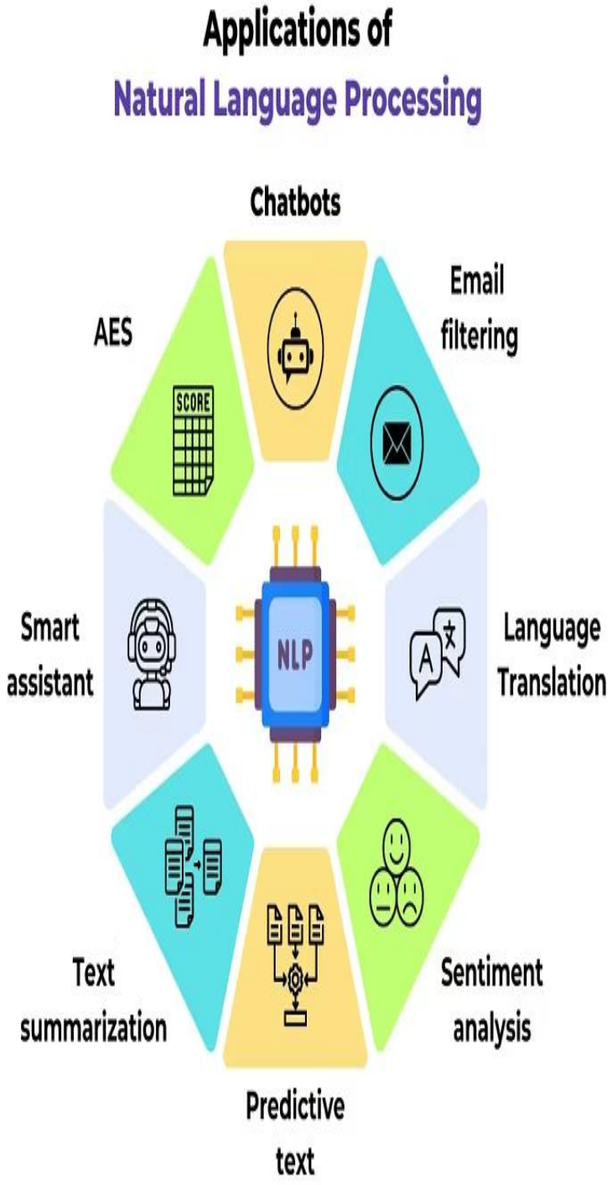
- Training accuracy keeps improving
- Validation accuracy **stagnates or decreases**
- Indicates the model is memorizing training data

Underfitting:

- Both training and validation accuracy remain **low**
- Indicates the model is too simple or not trained enough

Training vs Validation Graph





Results & Future Vision

Summary of Past Work

Consolidated findings demonstrate improvements in text preprocessing and embedding methodologies for NLP efficacy.

Impactful Results

Advanced preprocessing and embedding approaches lead to higher accuracy and faster convergence in classification models.

Future Directions

Focus on transfer learning for low-resource languages, ethical model explainability, and broader NLP application domains.

Future Directions

Model Enhancement

Explore more sophisticated NLP architectures and domain adaptation for improved detection accuracy.

Multi-Lingual Support

Extend models to identify hate speech across diverse languages and cultural contexts worldwide.

Real-Time Moderation

Implement systems capable of detecting and intervening against hate speech instantly on social platforms.

Collaboration

Encourage partnerships between researchers, platforms, and policymakers to combat online hate collectively.

Conclusion

Achievements

Developed models achieving high accuracy in detecting hate speech with practical real-world applications.

Ethical Balance

Addressing the challenge of protecting free speech while ensuring safety and inclusivity online.

Limitations

Current models face challenges like bias and language nuances; continued research is essential for improvement.


THANK YOU