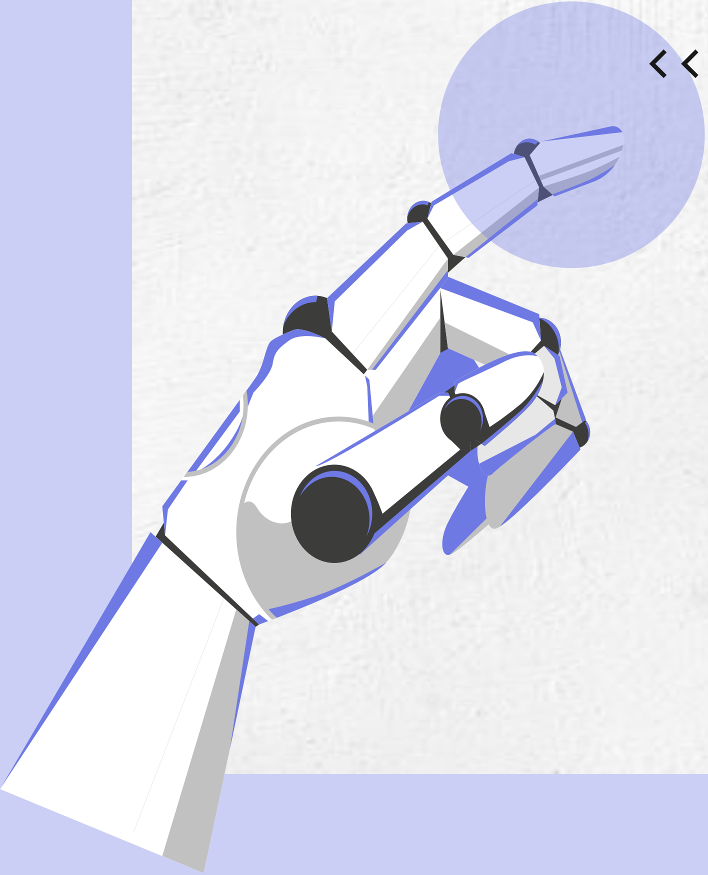


LITERATURE SURVEY ON APPLICATION AND DEVELOPMENT TREND OF N-GRAM MODEL IN NLP - CS6120

GROUP 6: ZHANSHUO DAI,
HIRANMAI DEVARASETTY
KRISHNA PRIYA GITALAXMI
SNEHITHA BARUKULA

INTRODUCTION TO THE N-GRAM MODELS & RESEARCH METHODOLOGY

What is a N-gram model?

•N-gram models are a fundamental technique in NLP that analyze the probability of word/character sequences.

•An N-gram is a contiguous sequence of n words or characters extracted from a given corpus. Common n-grams include unigrams(n=1), bigrams(n-2), and trigrams(n=3).These capture the statistical properties of the word sequences, providing a foundation for various NLP tasks.

Our Research:

•It involves a strategic search using IEEE Xplore and Google Scholar databases. We selected papers that were relevant to N-gram models in NLP, published within the last five years which are featured in high-impact journals and conferences, and frequently cited.

THEMATIC ANALYSIS OF THE N-GRAM MODELS

The Research papers on N-grams mainly covered the following themes/domains :

- Education Technology (EdTech)
- Information Technology
- Healthcare/Bioinformatics
- CyberSecurity

N-gram applications identified(most common-least common):

1. Text Classification
2. Information Retrieval
3. Language Modeling, Spell Correction, Sentiment Analysis
4. Text Generation,Text Compression,Machine Translation

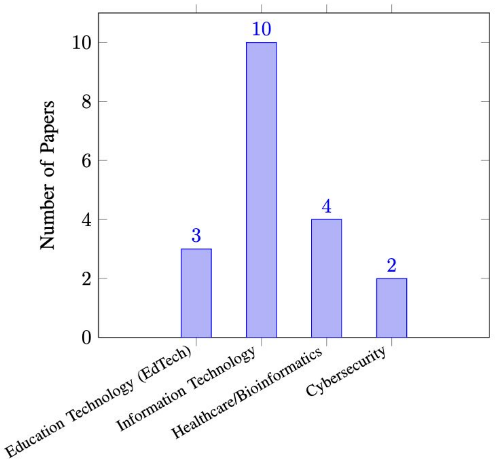


Fig. 2. Thematic Analysis of N-gram Model Research

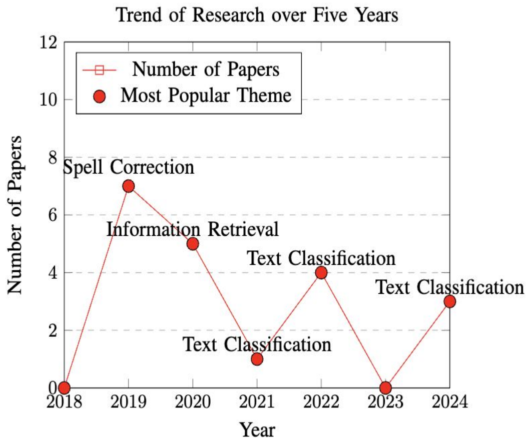


Fig. 3. Research Papers and Most Popular Themes over Five Years

METHODOLOGIES & ALGORITHMS

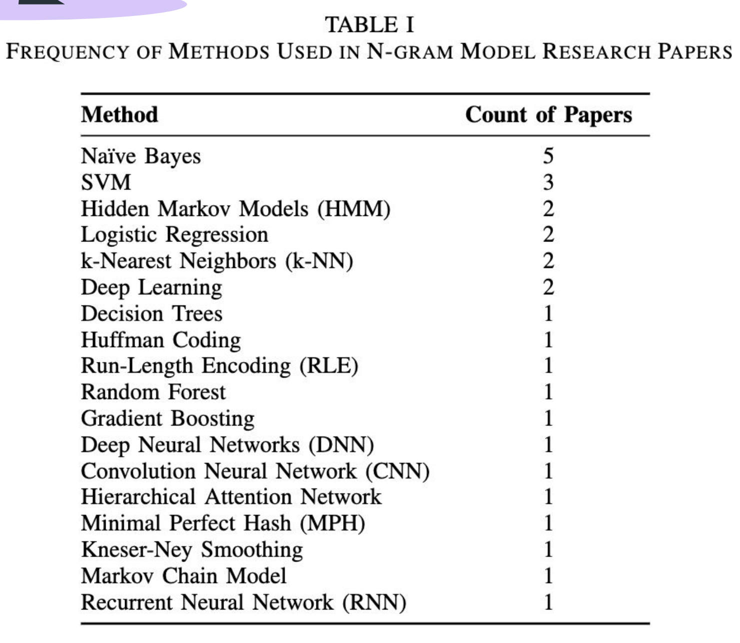


TABLE I
FREQUENCY OF METHODS USED IN N-GRAM MODEL RESEARCH PAPERS

Method	Count of Papers
Naïve Bayes	5
SVM	3
Hidden Markov Models (HMM)	2
Logistic Regression	2
k-Nearest Neighbors (k-NN)	2
Deep Learning	2
Decision Trees	1
Huffman Coding	1
Run-Length Encoding (RLE)	1
Random Forest	1
Gradient Boosting	1
Deep Neural Networks (DNN)	1
Convolution Neural Network (CNN)	1
Hierarchical Attention Network	1
Minimal Perfect Hash (MPH)	1
Kneser-Ney Smoothing	1
Markov Chain Model	1
Recurrent Neural Network (RNN)	1

The frequency of methods used in N-gram model research papers highlights the dominance of traditional techniques like Naïve Bayes and SVM, particularly in the most common application of N-grams: text classification. These methods are well-suited for categorizing text data based on N-gram features. Information retrieval also benefits from algorithms like Hidden Markov Models (HMM) and k-Nearest Neighbors (k-NN), which are effective in matching query terms to documents. Language modeling and spell correction often rely on methods like Logistic Regression and Decision Trees, which can predict the probability of word sequences and correct errors in text. Sentiment analysis, though increasingly using Deep Learning, still sees the application of these traditional methods to capture the sentiment expressed in text. Less frequently, N-grams are applied in text generation, text compression, and machine translation, where advanced methods like Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) are gaining traction, highlighting the evolving nature of research in these areas.

HANDLING OF N-GRAM DATASETS

Diverse Dataset Utilization : The research utilizes a wide range of datasets from social media posts to clinical texts, highlighting the flexibility and adaptability of N-gram models across different domains.

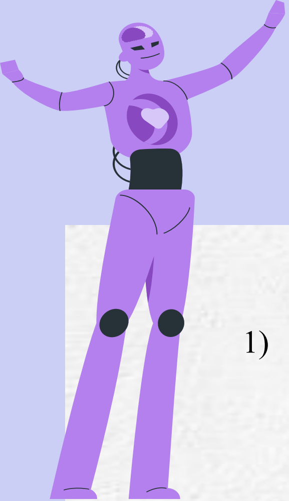
Language-Specific vs. General Applications

Examples of Dataset Applications

- **Chatbot Interactions**
- **E-Learning Reviews**
- **Software Code**
- **Entertainment and Media Reviews**
  a. **MovieLens and Yelp**
  b. **MR2004 Movie Reviews**

Impact of Dataset Choice on Model Design

- **Influence on Model Design and Eval** The choice of dataset significantly impacts how models are structured and evaluated, with different data types requiring specific features and processing capabilities.
- **Adaptability of N-Gram Models**: The wide application range from text compression to predictive analytics demonstrates the robustness of N-gram models to adapt to various data characteristics and requirements.



KEY CONTRIBUTIONS IN SENTIMENT ANALYSIS AND TEXT CLASSIFICATIONS

1) **Core Contributions of N-Gram Models :** Among the 20 articles reviewed, half explore the application layer of N-grams, with a significant focus (6 out of 10) on text classification, demonstrating the effectiveness of N-grams in this domain.
2) **Hybrid Models for Enhanced Accuracy**: Political Sentiment Analysis, Spam Detection
3) **Application in User Interaction**: Chatbot Enhancement, E-Learning Feedback Analysis
4) **Emerging Techniques and Privacy Enhancements**
   a. **Innovative Use in Localized Language Processing**:Hindi Language Text Generation
   b. **Federated Learning**: Decentralized Data Training

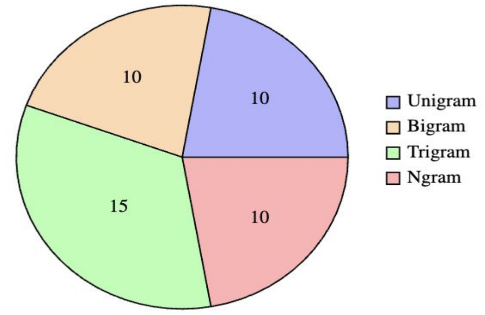
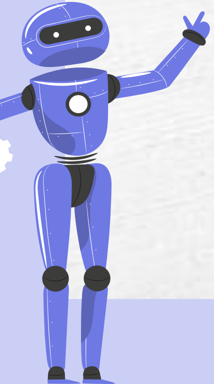


Fig. 4. Usage of Unigram, Bigram, and Trigram Models in Research Papers

NLP APPLICATIONS



The integration of n-gram models with deep machine learning techniques enhanced many applications.

Dynamic N-Gram System Based on an Online Croatian Spell Checking Service - Gledec

N-gram models for Text Generation in Hindi Language - Ghude

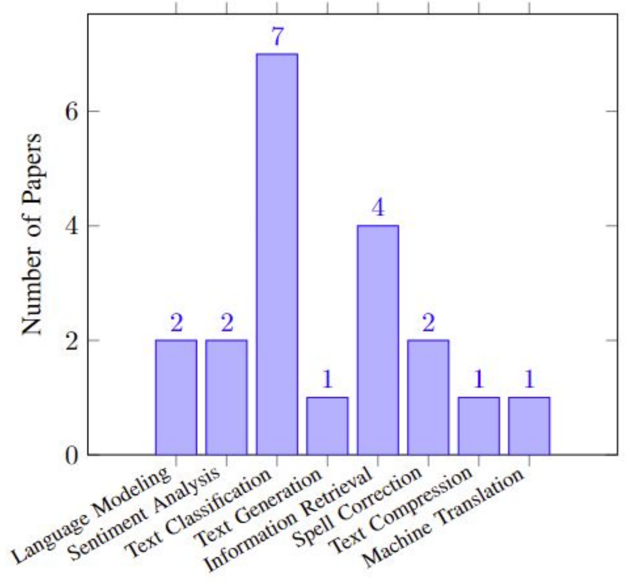


Fig. 1. Applications of N-gram Model across papers

COMPARATIVE ANALYSIS

- **Methodologies**
- **Datasets**
  a. social media data
  b. clinical text data
  c. chatbot interaction logs
  d. movie reviews dataset

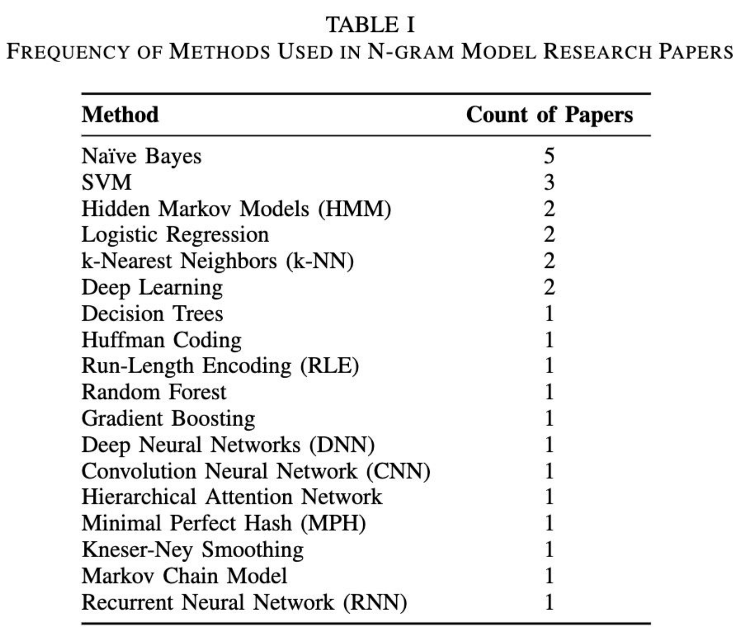
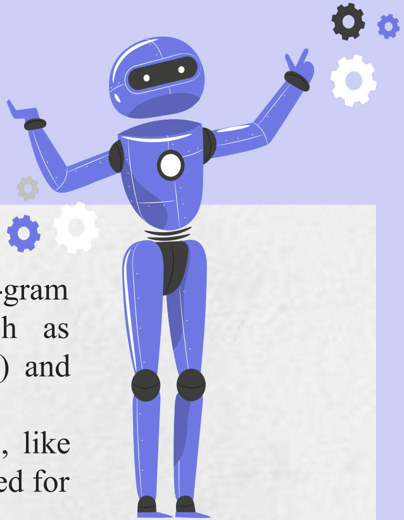


TABLE I
FREQUENCY OF METHODS USED IN N-GRAM MODEL RESEARCH PAPERS

Method	**Count of Papers**
Naïve Bayes	5
SVM	3
Hidden Markov Models (HMM)	2
Logistic Regression	2
k-Nearest Neighbors (k-NN)	2
Deep Learning	2
Decision Trees	1
Huffman Coding	1
Run-Length Encoding (RLE)	1
Random Forest	1
Gradient Boosting	1
Deep Neural Networks (DNN)	1
Convolution Neural Network (CNN)	1
Hierarchical Attention Network	1
Minimal Perfect Hash (MPH)	1
Kneser-Ney Smoothing	1
Markov Chain Model	1
Recurrent Neural Network (RNN)	1

PERFORMANCE EVALUATION



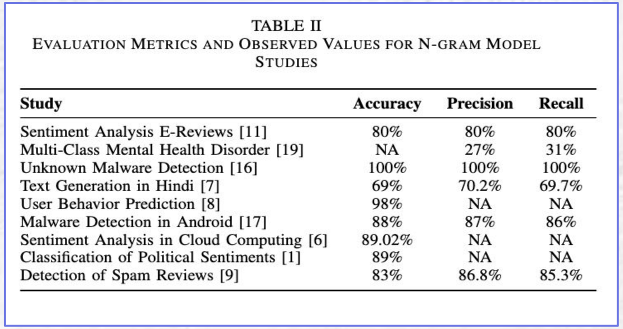


TABLE II
Evaluation Metrics and Observed Values for N-gram Model Studies

Study	Accuracy	Precision	Recall
Sentiment Analysis E-Reviews [11]	80%	80%	80%
Multi-Class Mental Health Disorder [19]	NA	27%	31%
Unknown Malware Detection [16]	100%	100%	100%
Text Generation in Hindi [7]	69%	70.2%	69.7%
User Behavior Prediction [8]	98%	NA	NA
Malware Detection in Android [17]	88%	87%	86%
Sentiment Analysis in Cloud Computing [6]	89.02%	NA	NA
Classification of Political Sentiments [1]	89%	NA	NA
Detection of Spam Reviews [9]	83%	86.8%	85.3%

- Table II shows the strong performance of N-gram models in specific NLP applications, such as Unknown Malware Detection(accuracy 100%) and sentiment analysis(accuracy 89.02%).
- Lower precision and Recall in complex tasks, like multi-class mental disorder, underscores the need for improved approaches in nuanced applications.

- Table III's F1 scores reveal that the N-gram models excel in certain applications but struggle in others, such as the multi-class mental health disorder with an F1-score of 20%.
- This disparity highlights the model's strength in well-defined tasks, while more complex tasks may need advanced ML techniques for better performance.

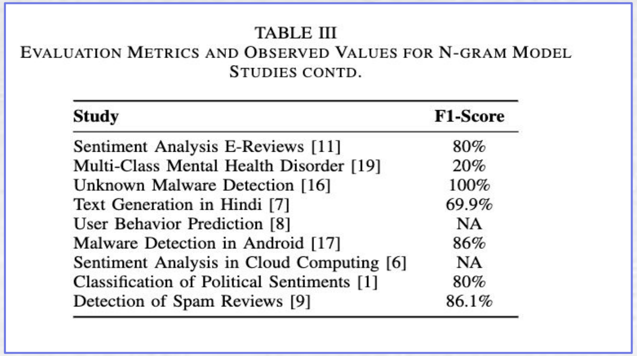


TABLE III
Evaluation Metrics and Observed Values for N-gram Model Studies contd.

Study	F1-Score
Sentiment Analysis E-Reviews [11]	80%
Multi-Class Mental Health Disorder [19]	20%
Unknown Malware Detection [16]	100%
Text Generation in Hindi [7]	69.9%
User Behavior Prediction [8]	NA
Malware Detection in Android [17]	86%
Sentiment Analysis in Cloud Computing [6]	NA
Classification of Political Sentiments [1]	80%
Detection of Spam Reviews [9]	86.1%

CONCLUSION & REFLECTION

- The literature review highlights the adaptability of the N-gram models in NLP, especially when integrated with other complex Machine Learning Techniques. However, the analysis of research papers highlights key limitations such as scalability, adaptability, dataset diversity, and platform scope.
- Addressing these challenges is essential to enhance the effectiveness and broaden the applicability of N-gram models, especially in complex NLP tasks and minority languages.
- Future research should focus on optimizing quantum-enhanced N-gram models for more robust and scalable solutions.

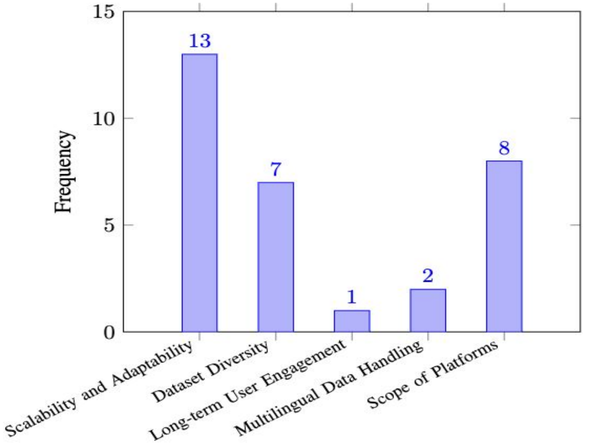
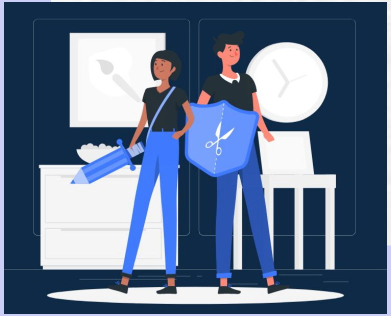


Fig. 5. Frequency of Limitation Categories in Research Papers



Reflection:

Our literature survey was a valuable learning experience in teamwork and problem-solving. Despite various challenges such as time conflicts, we collaborated effectively, refining our paper's structure and content.

THANK YOU!