

Journal Pre-proof

Unmasking Mnemonics – Leveraging Content Moderation Model for Decoding Encoded Communication in Digital Conversations

Sumithra S and Sujatha P

DOI: 10.53759/7669/jmc202505178

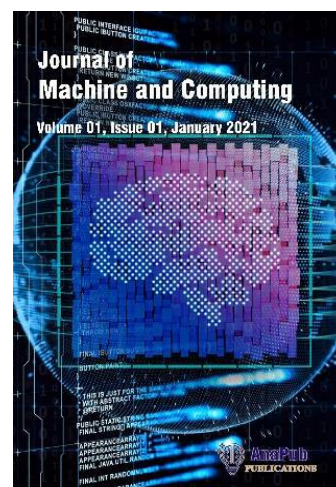
Reference: JMC202505178

Journal: Journal of Machine and Computing.

Received 12 April 2025

Revised from 22 May 2025

Accepted 29 July 2025



Please cite this article as: Sumithra S and Sujatha P, “Unmasking Mnemonics – Leveraging Content Moderation Model for Decoding Encoded Communication in Digital Conversations”, Journal of Machine and Computing. (2025). Doi: <https://doi.org/10.53759/7669/jmc202505178>.

This PDF file contains an article that has undergone certain improvements after acceptance. These enhancements include the addition of a cover page, metadata, and formatting changes aimed at enhancing readability. However, it is important to note that this version is not considered the final authoritative version of the article.

Prior to its official publication, this version will undergo further stages of refinement, such as copyediting, typesetting, and comprehensive review. These processes are implemented to ensure the article's final form is of the highest quality. The purpose of sharing this version is to offer early visibility of the article's content to readers.

Please be aware that throughout the production process, it is possible that errors or discrepancies may be identified, which could impact the content. Additionally, all legal disclaimers applicable to the journal remain in effect.

© 2025 Published by AnaPub Publications.



Unmasking Mnemonics – Leveraging Content Moderation Model for Decoding Encoded Communication in Digital Conversations

S.Sumithra¹, Dr.P.Sujatha²

¹ Research Scholar, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, India. sumithra99@gmail.com

² Professor and Head Department of Computer Applications, Vels Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai, India. suja.research@gmail.com

Abstract

In today's world, digital platforms have witnessed an explosion in the digital conversations and are not straightforward. A significant contributor to this complexity is the use of subtle references to another context or with encoded texts. These are said to be Mnemonics appearing in the form of Abbreviations, Numeronyms, Symbolic representations, Emoji-based codes, Leetspeak etc., in everyday communication. There are various types of mnemonics used in online conversations, which include phonetic substitutions (eg. Gr8 for 'great'), numerical encoding (e.g. 143 for 'I love you'), and symbolic representations (with emojis and icons), abbreviations (LOL for Laugh Out Loud) etc., This linguistic creativity is not only a tool for memory and efficiency, but also a growing challenge for automated moderation and content understanding systems, as mnemonics often encode non-explicit, sensitive, or policy-relevant meanings that typical keyword-based approaches might fail to identify. To address this gap, we introduce a Content Moderation Model, which is a large language model (LLM) based pipeline that systematically detects, categorizes, and deciphers both general and context-specific mnemonic constructs within user-generated text. This methodology builds upon advances in deep learning, leveraging the representational power and semantic flexibility of models such as GPT-4.1, known for their success in complex linguistic and content analysis tasks across domains. This framework uses a corpus of both harmless and sexually-coded user-generated texts to identify mnemonic patterns such as Phonetic substitutions, Emoji usage, and Leetspeak. The system accurately flags and classifies mnemonic types, enabling improved moderation, linguistic analysis, and platform policy design. The outcome—quantified through rigorous empirical validation, demonstrates substantial improvements in identifying and decoding diverse mnemonic forms. These findings provide actionable insights for platform policy, and the design of more accessible, inclusive communication systems that acknowledge both the benefits and risks of mnemonic language.

Keywords

Numeronyms, Emojis, Mnemonics, Phonetic Substitutions, Abbreviations, Leetspeak, Large Language Models, Content Moderation Model, Prompting, Zero shot Learning, FewShot Learning

1. Introduction:

Mnemonics, traditionally was evolved as a memory aid, to easily remember and recollect from memory. Social Media content, once straightforward, is increasing in its opacity and often includes cryptic symbols, abbreviated phrases, and disguised meanings. Whether chatting with friends, participating in online communities, or navigating social media, people increasingly rely on mnemonics: creative shortcuts like abbreviations, numeronyms, emojis, and leetspeak. These forms of encoded language do

more than just save time; they reflect our need for creativity, privacy, and sometimes, the desire to communicate beneath the surface of public scrutiny. This evolution is both fascinating and challenging. What began as simple memory aids has transformed into a complex web of symbols and codes, constantly evolving to keep pace with new platforms, cultural trends with increased anonymity, suppressing any unparliamentary words intentionally with indirect communication and escaping any policy restrictions in any social media platform. This linguistic creativity presents a significant challenge for online safety and moderation. Traditional content moderation tools, built on straightforward keyword detection or basic machine learning, often falter when faced with the ingenuity of modern mnemonics. As a result, harmful or sensitive content can slip through the cracks, while benign messages are sometimes unfairly flagged. This paper seeks to bridge that gap. By harnessing the power of large language models (LLMs), to decode and classify the diverse landscape of mnemonics in digital conversations. Our goal is not just technical accuracy, but also a deeper understanding of how people communicate in the digital age—balancing the need for safety with respect for creativity and cultural nuance. Through this research, we hope to empower platforms, moderators, and researchers to foster safer, more inclusive, and more expressive online spaces.

2. Literature Review:

In [1] Stone, C.B et.al., mentioned that mnemonic effects of social media use, with 90% of American adolescents and 65% of adults actively engaging online. It explores how information type (personal vs. public) and user roles (producer vs. consumer) influence memory, highlighting induced forgetting, false memories, and truthiness, while identifying key areas for further research. In [2] Jaewook Lee et.al., mentioned that mnemonic vocabulary is underexplored, and explored automating keyword mnemonics for vocabulary learning using an overgenerate-and-rank method with large language models (LLMs). By generating and evaluating verbal cues through psycholinguistic metrics and user studies, the authors find that LLM-generated mnemonics rival human ones in quality, though learner preferences vary widely. But he did not explore more on identifying mnemonics in the existing social media. In [3] Roediger, H. L. mentioned about 4 mnemonic methods – a. Imagery, for forming mental images to remember words, b. Link Method – creating associations between items in a sequence through visual or narrative links, c. Peg System – associating items with a pre-memorized list of “pegs” (e.g., rhyming words or numbers), d. Loci method – placing items to be remembered along a familiar mental route or location sequence. In [4], Gupta et.al., analyses pedophile chat conversations, using online grooming theory and perform a series of linguistic-based empirical analysis on several pedophile chat conversations to gain new insights and patterns. In [8], Satadruta Mookherjee et.al., explores how consumer loneliness influences preference for mnemonic features in social media, affecting platform choice and consumer behavior and how consumers prefer SnapChat which is linked to mnemonics unlike Facebook consumers. In [9] Ana Lúcia Migowski da Silva, says how social media, specifically Facebook, shapes memories and discourses about Brazil's dictatorship (1964-1985) through technological and mnemonic practices. In [10], Barber et.al., examines how age, emotion, and social context influence retrieval-induced forgetting (RIF). Both individual (WI-RIF) and socially shared (SS-RIF) forgetting occurred equally across age groups and emotional content. However, SS-RIF only emerged when listeners heard from same-sex speakers, suggesting people are more likely to corrective—and thus forget related information—with those they feel closer to, impacting both personal and collective memory. In [11], Obiora et.al., examines, how university students in Anambra University in Nigeria uses emojis for emotional expression and communication on social media. With 72% of respondents able to decode common emojis, the findings highlight emojis' crucial role in enhancing social interaction. The study recommends integrating emojis into broader communication practices for efficiency and clarity.

3. Evolution of Mnemonics in Digital Contexts:

Mnemonics at its core, originated as memory aids, acts as tools that help individuals to encode, retain and recall information efficiently. These include a variety of cognitive strategies such as:

Keyword Mnemonics: Helps in associating foreign language vocabulary with vivid imagery(e.g, remembering the Spanish word 'gato' for cat by picturing a cat of the gate).

Chunking: Grouping information, such as splitting a phone number into smaller segments to make the pattern more memorable.

Musical Mnemonics: Using songs or rhythms, like the ABC song, to reinforce sequences.

Acronyms and Acrostics: Creating new words or phrases from the first letters of a series (e.g., "LOL" for "Laugh Out Loud," or acrostics like "LOVE" for "Lasting Connection beyond words. Overcomes obstacles with grace, Values each moment shared, Elevates the soul through affection").

Rhymes and Connections: Employing rhyme and association to make recall easier and more enjoyable.

Method of Loci and Peg Methods: Placing items along a mental route or associating them with a pre-memorized list (e.g., "1-thumb, 2-shoe, 3-tree...").

Link Method: Creating stories or visual images that connect items to be remembered.

3.1 From Memory Aid to Digital Code

With the rise of digital communication, mnemonics have undergone a significant transformation. No longer limited to memory enhancement, they now serve as tools for brevity, creativity, privacy, cultural expression, and, increasingly, moderation avoidance in online interactions. The shift from traditional to digital mnemonics has introduced new forms that are now integral to online conversations:

- **Abbreviations:** Shortened expressions (e.g., "OMW" for "On My Way").
- **Numeronyms:** Numbers substituting for words or syllables (e.g., "143" for "I Love You," "I18n" for "Internationalization").
- **Phonetic Substitutions:** Using numbers or letters that sound like parts of words (e.g., "Gr8" for "Great," "8r" for "Later").
- **Emoji Encoding:** Emojis representing emotions, objects, or even complex ideas (e.g., 😊 for smile, 🍆 for penis in adult contexts).
- **Leetspeak:** Letters replaced with numbers or symbols to evade moderation or create in-group language (e.g., "N00b" for "Newbie," "S3x" for "sex").

Slang and Euphemisms: Informal or coded language for cultural or privacy reasons (e.g., "bae" for "babe" or "before anyone else," "thicc" for "curvy").

3.2 Summary of Digital Mnemonics:

| Type | Example | Context | Motivation |
|--------------|---------|-----------------|-------------------------------|
| Abbreviation | LOL | General | Brevity, Expression |
| Numeronym | 143,69 | Romantic, Adult | Privacy, Moderation Avoidance |

| | | | |
|-----------|------------|-------------------|----------------------|
| Phonetic | Gr8, L8r | General | Brevity, Creativity |
| Emoji | 🍆, 💦 | Emotional, Adult | Expression, Privacy |
| Leetspeak | S3x, N00b | Adult | Moderation Avoidance |
| Slang | Bae, thicc | Youth, subculture | Identity, Creativity |

Table [1] – Summary of Digital Mnemonics

3.3 Contextual and Sensitive Usage:

The use of mnemonics in digital conversations is often context-dependent. For example, in adult or sensitive discussions, numeronyms like "69" or emojis like "🍆" and "💦" are used to convey sexual meaning while evading explicit language filters. Similarly, leetspeak and creative spelling (e.g., 1-0-0# for "fuck") help users bypass automated moderation systems.

4. Overview of Earlier Content Moderation Approaches:

4.1 Traditional Machine Learning Models (SVM, Random Forest)

Early attempt on Content Moderation includes a traditional Keyword search of sexual or abusive words, or a deployment of traditional Machine Learning Models such as Support Vector Machine (SVM), Random Forests. These models traditionally use features like n-grams, Bag of Words (BOW) or TF-IDF vectorizer to represent text. While computationally efficient, these models struggle with the nuanced context-dependent nature of mnemonics as the language evolves rapidly.

Support Vector Machine Mechanism for Multi-class Classification:

One-Vs-Rest(OVR):

- For k classes, train k binary classifiers.
- During Prediction, for input x, all classifiers are evaluated, and the one with the highest output decision value is chosen.

One-Vs-One(OvO):

- For k classes, train $k(k-1)/2$ binary classifiers.
- Each classifier (i,j) distinguishes between classes i and j. The input is assigned to the class with the most "wins" in pairwise contests.

Random Forest Mechanism for Multi-class Classification:

- Each decision tree in the forest predicts a class label or the input.
- For multi-class classification, the final output will be the class with the majority vote across all the trees.

$$\text{Prediction}(x) = \arg \max_c \sum_{t=1}^N 1(\text{Tree}_t(x) = c)$$

4.2 BERT and Transformer-Based Models

The introduction of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), marked a significant advancement. BERT's contextual embeddings allow for a deeper understanding of word meaning based on surrounding text, improving detection of abbreviations, numeronyms, and some forms of leetspeak. However, more contextual text or highly creative texts, will still be failed as it requires such texts in training data to eventually understand.

5. Limitations of Existing Moderation and Detection Systems:

Content Moderation Systems face a lot of challenges in detecting Mnemonics. It is due to coded language, symbolic references, emojis. Some Mnemonics even represent double meaning words, and some are highly contextual. They convey sensitive or prohibited content while evading the moderation algorithms. With just NLP-based Machine Learning Techniques, the model tends to lose the context, and the entire conversation may not be understood correctly.

Contextual Understanding Issues:

Moderation tools that use machine learning models, often struggle to understand the complete context, making it difficult to discern the intended meaning behind the mnemonic. Most of the mnemonics used in adult conversations are contextual and not explicit.

Rapid Evolution of Language and Slang:

Online conversations frequently change or develop new slang to communicate. They use mnemonics, emojis, abbreviations etc., to communicate. This dynamic evolution outgrows the ability of the Machine Learning Models. Continuous update to the model is required and is highly expensive, leading to gaps in the moderation.

Multilingual and Dialectical Variation

Often, Moderation Systems are trained on English language data, limiting their understanding to multilingual conversations or multiple blended language conversations. For example, mixing up adult content words in 2 different languages blended in a sentence. These limitations hinder the detection of mnemonics in non-English contexts.

High False Positive and Negative Rates:

The complexity of Mnemonics contributes to higher rates of false positives and false negatives leading to distrust in the effectiveness of moderation systems.

Motivation behind Mnemonic Usage:

Main intention or motivation behind the mnemonic usage is to avoid Moderation systems, expressing Cultural motive, Real-Time Expressions of feelings and emotions, Anonymity and Privacy.

Moderation Avoidance:

Current Social Media platform includes either Human content Moderators or Moderator System. Moderator systems were trained predominantly with Machine Learning Models which helps to filter the abusive or sexual words or alert the user as a proper moderation system. But these systems fail to detect the words from growing mnemonical vocabulary, as these mnemonics vary from country to country as well. Moderation Avoidance happens in communities where discussion of sexuality or identity are subjected to automatic suppression.

Cultural Expression:

Mnemonics are used by subcultural norms. Emojis such as ‘Eggplant’ or coded words like ‘Alphabet Mafia’ (LGBTQIA+) develop into community specific mnemonics.

Real-Time Expressions:

Chat apps such as Whatsapp, Snapchat allows mnemonics, emojis for users to express their feelings, complex emotions and intentions with a more simplified way with just few characters.

Anonymity and Privacy:

In Social media platforms mnemonics provide a protective linguistic layer. They allow the discussion of personal, intimate or socially stigmatized content without direct exposure.

Code-Switching in Youthful Digital Flirtations:

In some chats, young adults and teenagers often engage with in subtle forms of code-switching to express romantic, sext, and flirtatious interest. This involves mnemonics, abbreviations, emojis, and symbolic language that convey deeper meaning without obvious declaration. Such linguistic creativity allows individuals to flirt playfully while escaping the boundaries of social norms and digital platform guidelines.

5.1 Ethical and Legal Considerations:

- **Child Safety Concerns:** Mnemonics can easily bypass filters that are designed to protect children
- **Freedom of Speech vs Regulation:** Balancing expression and platform responsibility is challenging.
- **Bias in Moderation Models:** Models might overflag marginalized communities using reclaimed or coded language.

6. Approach & Methodology:

A set of data from various social media platforms has been gathered and it is pre-processed. The cleaned and pre-processed data is sent to various models (SVM, Random Forest, BERT) and to the Content Moderation Model (CMM) to find the best model for Mnemonics detection and its accuracy.

6.1 Usage of Large Language Models:

Recent advancements in Large Language Models, such as Gpt-4.1 have pushed their boundaries of mnemonics detection, as it has been trained on massive and diverse corpus. LLMs can decode context-dependent and different forms of mnemonics, including those phonetic, symbolic and cultural elements. Their generative and few-shot capabilities make them well-suited for nuanced moderation and for adapting to the ever-changing digital communications. These large Language Models can further be fine-tuned to suit the custom needs on top of the existing knowledge.

6.2 Dataset Annotations and Statistics:

6.2.1 Annotation Process:

Each sentence in the dataset was manually reviewed and labelled. The annotation focused on identifying the presence and type of mnemonics, with the special attention to distinguish between general and sexual mnemonics.

Labelling Schema:

| Labelling Schema | Description |
|---------------------|--|
| Mnemonic Type | <ul style="list-style-type: none">• Abbreviation,• Leetspeak,• Slang,• Numeronym,• Phonetic Substitution,• Euphemism,• Emoji, Symbolic Representation• None |
| Moderation Category | <ul style="list-style-type: none">• General• Sexual• None |

Table [2] – Labelling Schema

6.2.2 Dataset Statistics:

Total Size:

General Conversations (No Mnemonics): 300 sentences

General Conversations (With Mnemonics): 300 sentences

Sexual Conversations (With Mnemonics): 1000 sentences

Dataset Samples:

General Conversations without any Mnemonics:

| Sentence | Type | Platform | Notes |
|-----------------------------------|------|----------|------------------|
| Are you free this evening? | None | Discord | No mnemonic used |
| This blanket is so cozy. | None | YouTube | No mnemonic used |
| I lost my keys this morning. | None | TikTok | No mnemonic used |
| I need to finish this assignment. | None | Discord | No mnemonic used |

Table [3] – General Conversations without any Mnemonics

General Conversations with Mnemonics:

| Sentence | Type | Platform | Notes |
|--|--------------|-----------|---|
| BRB, making some n00dles | Leetspeak | Reddit | 'N00dles' = 'noodles', fun usage |
| OMW to the gym | Abbreviation | Instagram | 'OMW' = on my way |
| She gave off major main character energy | Slang | TikTok | 'Main character energy' = confidence/aura |
| OMW to the gym | Abbreviation | Reddit | 'OMW' = on my way |

Table [4] – General Conversations with Mnemonics

Sexual Conversations with Mnemonics:

| Sentence | Type | Platform | Notes |
|---|-----------------------|-----------|---|
| She asked for lewds right off the bat before the morning meeting. | Abbreviation | Discord | 'Lewds' = nude photos |
| He invited me over for a nightcap... yeah right at 2am. | Euphemism | Instagram | 'Nightcap' used as coded invite for sex |
| Send noods, not moods at 2am. | Phonetic Substitution | Discord | 'Nudes' → 'Noods' |
| He typed 'sh3 luvs 2 rid3 it raw'... wild. | Leetspeak | Discord | 'Sh3' = 'She', 'rid3' = 'ride' |
| He's not just into cuddles, if you know what I mean | Euphemism | Instagram | Cuddles = sexual activity |

Table [5] – Sexual Conversations without any Mnemonics

6.3 Overall Architecture for Decoding Mnemonics:

Overall Architecture for decoding mnemonics is a three-step process, which includes Pre-Processing, Text Processing, and Post-Processing Layers. Each layer is explained in detail below. Also, it involves fine-tuning of a Large Language Model (GPT-4.1) as a Content Moderation Model.

6.3.1 Content Moderation Model:

For the Large Language Model to specifically act as a Content Moderation Model, the sexual, Mnemonic dataset has been collected, ingested for finetuning with appropriate results.

6.3.1.1 Content Moderation Model Objective:

The objective is to identify, classify texts as adult content, non-adult/generic content. Identify Mnemonic provided in the context and classify it as sexual/general. This model acts as a multi-class classifier.

6.3.1.2 Data Preparation:

- Data has been gathered from various social media platforms and labelled them
- Framed JSONL format with each line being a JSON object with Input(prompt) and Output(classification), ensuring data quality.

6.3.1.3 Upload Data, Configure, and Fine-Tune:

- Data was uploaded in OpenAI's platform, with GPT-4.1 as the base model.

- Created the fine-tuning job with the following hyperparameters.
 - Batch Size** - No. of Examples in each batch. As larger batch requires more memory and GPT, set the Batch size to 8.
 - Learning Rate** – Scaling factor for the learning rate. Set to low learning rate at 1, to avoid catastrophic forgetting and instability.
 - Epochs** – No. of Epochs to train the model. It was set to 3, to monitor validation loss or accuracy
- Job was started and tracked for its completion. Now, this becomes the Content Moderation Model for the rest of the architecture.
- The model was then deployed to integrate into the Mnemonic Decoder Architecture

6.3.2 Architecture for Decoding Mnemonics:

The below diagram explains how the data is pre-processed, and passed to LLM and how the data is post processed.

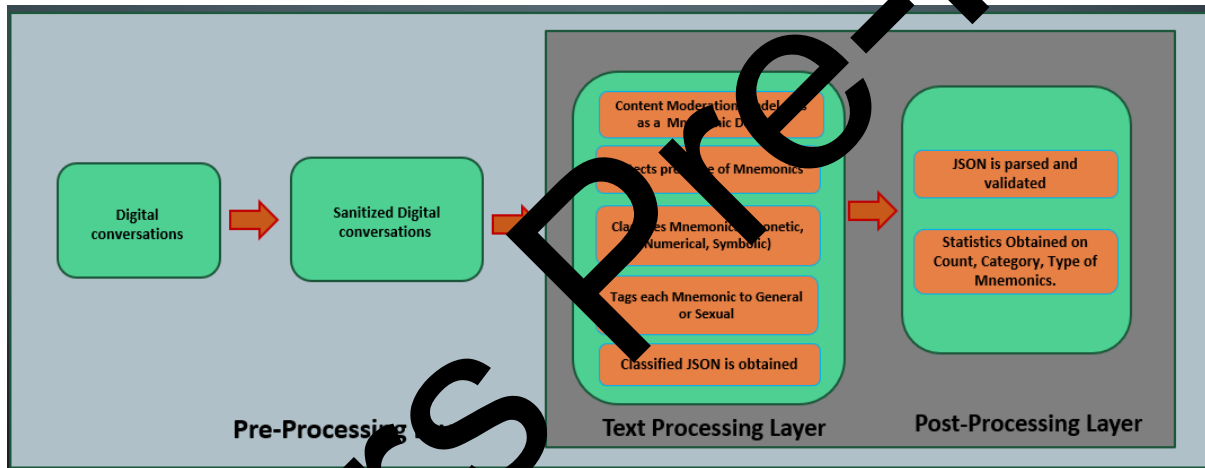
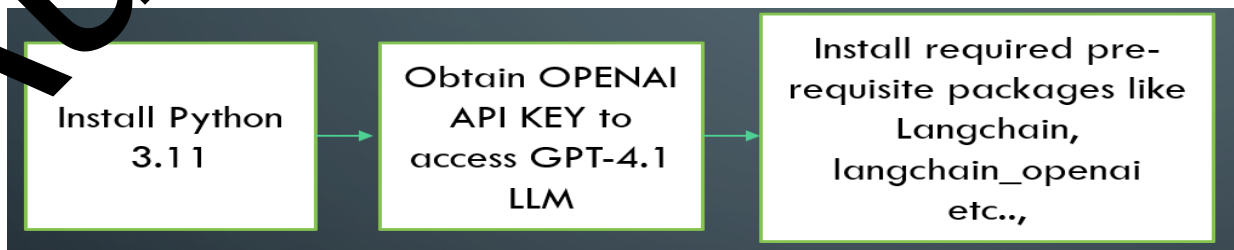


Fig [1] – Architecture for Decoding Mnemonics

6.3.3 Pre-requisites for Content Moderation Pipeline development:

To implement this Content Moderation Pipeline, a set of prerequisites has been done. An account has been created on OpenAI to generate the OpenAI API Key. The programming language has been set to Python 3.11, and the required packages such as openai, langchain_openai etc, have been installed for performing the below task. Large Language Model is set to the fine-tuned Content Moderation Model for detecting mnemonics and categorizing it as General Mnemonics or Sexual Mnemonics and the type of Mnemonics



Fig[2] – Pre-requisites

6.3.4 Pre-Processing Pipeline

Pre-processing includes 3 processes

- **Input:** Sentence from digital conversation (general, mnemonic-laced, or sexual).
- **Sanitization:** Basic cleaning and formatting to prepare the text.
- **Routing:** Text is passed to the Content Moderation Model based detection model.

6.3.5 Text Processing Pipeline:

LLM Invocation: Processed Sentence is sent to Content Moderation Model via a structured prompt.

Prompt Engineering for Content Moderation Model:

Prompt Engineering is essential for the Content Moderation Model. This helps to provide a Role, Instructions, and Examples for the model to perform its classification properly.

Prompt Design:

Role Definition:

The role is defined as "**Mnemonics Decoder**", giving the model a clear functional identity. This prompt role primes the model to focus on its task with purpose and clarity—an essential step in steering generative behaviour. By explicitly stating, "*You are a good Mnemonics Decoder,*" the prompt ensures that the model treats the task as one involving pattern recognition, semantic interpretation, and categorization, rather than casual language generation.

Few-Shot Examples in Prompt:

The prompt structure itself implicitly leverages few-shot principles by clearly defining categories (phonetic, numerical, symbolic), offering illustrative examples (e.g., "gr8," "143," "🍆"), and establishing expectations for structured output. These components serve as informal demonstrations or "shots" to guide the model's behavior.

The core logic was implemented using the Content Moderation Model with a temperature of 0.3. A structured prompt was designed to elicit detection and categorization of mnemonics.

6.3.6 Post Processing Layer for Content Moderation Model:

- **Parsing JSON Response** is parsed and validated.
- **Storage:** Results are added to respective datasets (CSV/Database).
- **Analytics Support:**
 - Flags content for moderation if any sexual mnemonics are detected.
 - Updates stats: count, category frequency, types of mnemonics used.

Output Structure and Formats:

Each sentence is analyzed and returns JSON containing:

- found_mnemonics (boolean)
- mnemonics_list: list of mnemonic items, each containing:

mnemonic, category, meaning, mnemonic_type

Following is the sample output received from the model.

Sentence: “*Send me those lewds*” – This is the input text to the model.

Detected: *True* – This states that Mnemonic has been identified in the input text.

Mnemonic: This gives the list of mnemonics identified, its category, meaning and the type of mnemonics. In our input,

Identified mnemonic is “*lewds*”. It is of type, “Phonetic Substitution”. This is the substitution for word “nudes”. It is of type “Sexual”.

Also there is one more Symbolic mnemonic, which represents flirtation and is also of type “Sexual”.

| Sentence | Detected | Mnemonic Identification |
|---------------------------|----------|--|
| Send me those lewds 👉👉 | True | "mnemonic": "lewds", "category": "Phonetic substitution", "meaning": "nudes", "mnemonic_type": "sexual" |
| | | "mnemonic": "👉👉", "category": "symbolic representation", "meaning": "flirtation", "mnemonic_type": "sexual" |

Table [6] – Mnemonic Identification Output

7. Statistical Evaluation Framework and Test Results

Decoding Mnemonics have been run across 4 models – SVM, Random Forest, BERT and Content Moderation Model.

7.1 Rationale for Model Selection

- **SVM and Random Forest:**

Provide interpretable baselines and highlight the limitations of feature-based methods in nuanced, context-dependent tasks.

- **BERT:**

Chosen for its strong contextual understanding and proven performance in text classification.

7.2 Training, Validation, and Test Splits:

With Respect to SVM, Random Forest and BERT, following are the Training, Validation and Test Splits.

| Split | Percentage | Stratified |
|------------|------------|------------|
| Training | 70% | Yes |
| Validation | 15% | Yes |
| Test | 15% | Yes |

Table [7] – Training, Validation, and Test Splits

7.3 Hyperparameter and Optimization:

For SVM, we performed grid search over kernel types (linear, RBF), C values (0.1, 1, 10), and gamma (0.01, 0.1, 1) using 5-fold cross-validation. For Random Forest, we tuned the number of estimators (100, 200), max depth (None, 10, 20), and criterion ('gini', 'entropy'). For BERT, the model was fine-tuned using the 'bert-base-uncased' checkpoint, with a learning rate of $2e-5$, batch size of 16, and 3 epochs. AdamW optimizer and early stopping on validation loss were used. All hyperparameters were selected based on validation performance.

For the Content Moderation Model, we have set the Batch Size to 8, Learning Rate to 1 and Epochs to 3

| Model | Key Hyperparameters Tuned | Optimization Method |
|--------------------------|--|---------------------|
| SVM | Kernel, C, gamma, class_weight | Grid Search |
| Random Forest | N_estimators, max_depth, criterion | Random search |
| BERT | Learning rate, batch size, epochs, max_len | Grid search |
| Content Moderation Model | Batch Size, Learning Rate, Epochs | Trial and Error |

Table[8] – Hyperparameter Configuration

7.4 Content Moderation Model Specific Parameters:

Temperature:

- The temperature parameter controls the output randomness in Content Moderation Model
- Sensitivity analysis is performed by evaluating at multiple temperatures, with observed impacts on consistency and accuracy.
- After the sensitivity analysis, the temperature was set at 0.3 where the accuracy was more

Prompt Design

The prompt includes role, role definition, and its task, and a few-shot examples. This is much required for the LLM to do only the designated task. When LLMs are provided with specific task instead of being generally invoked, they may not perform the required action.

Role and Role definition : “You are a Mnemonic Decoder...”

Few-shot Examples: To guide the model on how to detect various mnemonics

Prompt variations and temperature variation studies are reported, showing how changes affect the model performance in table below

7.5 Evaluation Framework:

For each model, the following statistical evaluation framework is employed to ensure robust, fair, and interpretable assessment of mnemonic detection performance.

Evaluation Metrics

Accuracy:

The proportion of total correct predictions (both positive and negative) overall prediction.

Precision, Recall, and F1-score:

- Precision: The fraction of true positive predictions among all positive predictions (model's exactness).
- Recall (Sensitivity): The fraction of true positives detected among all actual positives (model's completeness).
- F1-score: The harmonic mean of precision and recall, balancing both metrics.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad \text{----- (1)}$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad \text{----- (2)}$$

$$\text{F1} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}) \quad \text{----- (3)}$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad \text{----- (4)}$$

Confusion Matrix:

A table showing true positives, true negatives, false positives, and false negatives for each class, enabling detailed error analysis.

7.6 Statistical Significance Testing:

Paired t-tests:

Used to compare the performance (of F1-Scores) of different models across the same test samples, and determine if observed differences are statistically significant.

Chi-square test:

As our scenario is a kind of multi-class classification, Chi-square tests are applied to confusion matrices to test if the distribution of predicted vs actual classes differs significantly between models.

Similarity Co-efficient Evaluation:

Jaccard Similarity Index and Dice Co-efficients are used as a part of evaluation metrics for mnemonic detection across all models (SVM, Random Forest, BERT, LLM). These metrics are widely used in NLP to assess the similarity between sets, such as predicted labels vs actual labels, or between sets of detected mnemonics and ground truth.

Why Jaccard and Dice is required:

- These are used to compare the set of mnemonics detected by the models in each sentence to the set of mnemonics annotated as ground truth.
- In our case, a sentence contain multiple mnemonics, and these metrics provide a more nuanced measure than accuracy or F1-score alone
- Used for quantifying partial matches, especially when a model detects some, but not all relevant mnemonics.

Jaccard Similarity Index:

- Measures the size of the intersection divided by the size of the Union of two sets.
- Valid Ranges are from 0 (no overlap) to 1(identical sets)
- Used to compare the sets of predicted mnemonics to the ground truth set in text or label classification tasks

Formula

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Dice Co-efficient:

- Measures the similarity as twice the intersection divided by the sum of the sizes of the two sets.
- Valid Ranges are from 0(no overlap) to 1(identical sets)
- Used for evaluating string or set similarity and is closely related to the Jaccard Index.

Formula

$$D(A,B) = \frac{2 * |A \cap B|}{|A| + |B|}$$

7.7 Test Results

Model Accuracy with varied prompt and Temperature

| Model | Prompt Type | Temperature | Accuracy | F1-score | Notes |
|--------------------------|----------------|-------------|----------|----------|-----------------------|
| Content Moderation Model | Few-shot, role | 0.3 | 0.98 | 0.99 | Main results |
| Content Moderation Model | Zero-shot | 0.3 | 0.95 | 0.96 | For a variation study |
| Content Moderation Model | Few-shot, role | 0.7 | 0.96 | 0.97 | Higher Randomness. |

Table[9] - Model Accuracy with varied prompt and Temperature

Model Results with Chi-square and Paired t-test

| Model | Accuracy | Precision | Recall | F1-score | Chi-square(vs LLM) | Paired t-test (vs LLM) |
|--------------------------|----------|-----------|--------|----------|--------------------|------------------------|
| SVM | 0.70 | 0.68 | 0.66 | 0.67 | 25.8(p < 0.0001) | t=-6.2(p<0.0001) |
| Random Forest | 0.75 | 0.73 | 0.72 | 0.72 | 20.3(p<0.0001) | T=-5.7(p<0.001) |
| BERT | 0.85 | 0.84 | 0.83 | 0.83 | 9.1(p=0.003) | t=-3.4 (p=0.001) |
| Content Moderation Model | 0.98 | 1.0 | 0.98 | 0.99 | | |

Table[10] – Model Results with Chi-square and Paired t-test

Result Interpretation:

- Content Moderation Model has demonstrated the highest accuracy and F1-score, outperforming the other models across all metrics.
- Statistical Significance Test (chi-square and paired t-test) shows the performance difference between Content Moderation Model(LLM) and the baseline models.
- Precision and Recall increase consistently from SVM to Content Moderation Model(LLM) indicating improved detection and classification of mnemonics with more advanced models.

Model Results with Jaccard and Dice

| Model | Accuracy | Precision | Recall | F1-score | Jaccard | Dice |
|--------------------------|----------|-----------|--------|----------|---------|------|
| SVM | 0.70 | 0.68 | 0.66 | 0.67 | 0.51 | 0.62 |
| Random Forest | 0.75 | 0.73 | 0.72 | 0.72 | 0.57 | 0.66 |
| BERT | 0.85 | 0.84 | 0.83 | 0.83 | 0.71 | 0.77 |
| Content Moderation Model | 0.98 | 1.0 | 0.98 | 0.99 | 0.87 | 0.93 |

Table[11] -Model Results with Jaccard and Dice

Results Interpretation:

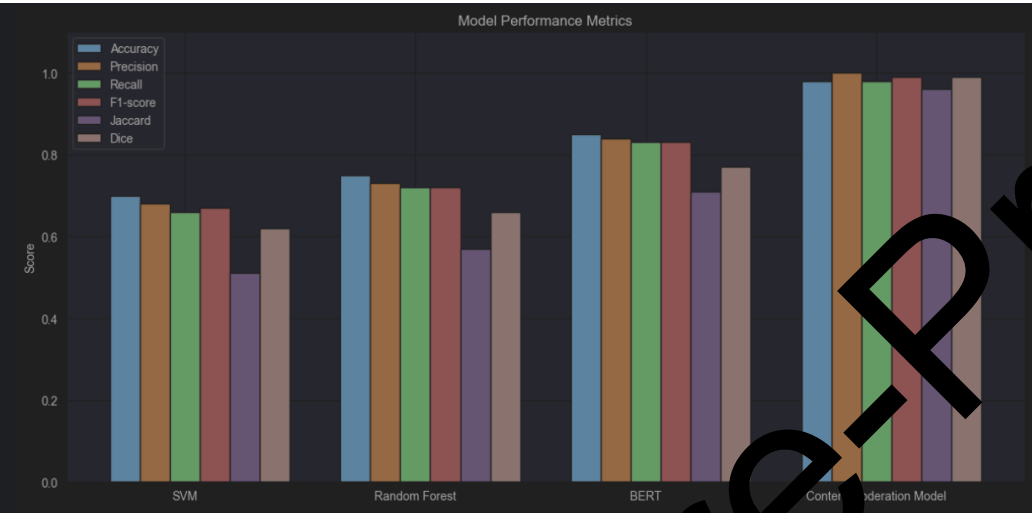
- Typically Jaccard score is lesser than the F1-score for same predictions in case of multi-class classification and Dice is closely related to F1-score and will be very similar in value, especially when classes are balanced.
- Note, Jaccard and Dice scores increase with model performance, reflecting better overlap between predicted and true mnemonics.

- Content Moderation Model(LLM) achieves the highest similarity and overlap, consistent with its superior accuracy and overlap.

7.8 Visualization for Test Results

The following are the various charts that support the results:

7.8.1 Performance Metrics across models



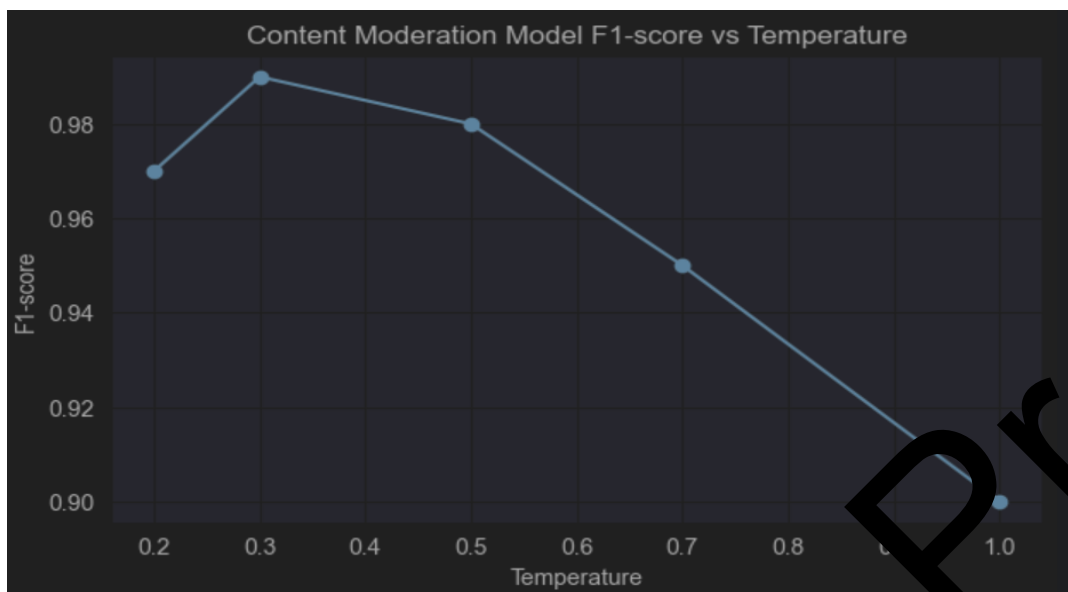
Fig[3] – Performance Metrics across models

7.8.2 F1-Score distribution across models:



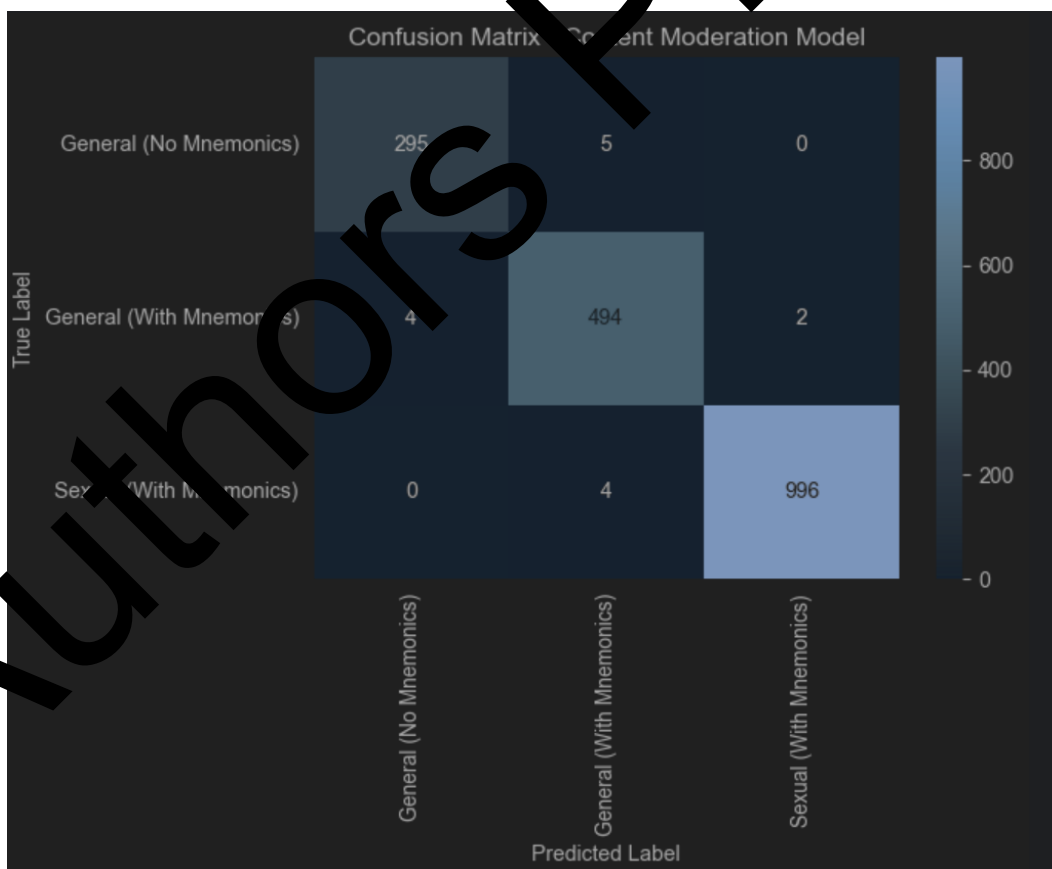
Fig[4] – F1-Score distribution across models

7.8.3 Content Moderation Model F1-Score vs Temperature



Fig[5] – Content Moderation Model F1-Score vs Temperature

7.8.4 Confusion Matrix for all Datasets



Fig[6] -Confusion Matrix for all Datasets

8. Conclusion:

This study demonstrates that Content Moderation Model is highly effective at detecting and categorizing mnemonics—including coded sexual expressions—in digital conversations. The model excels at recognizing a wide range of mnemonic patterns, such as phonetic substitutions, emoji encodings, and numerical codes, achieving near-perfect accuracy and F1-scores. This capability marks a substantial advance over traditional moderation systems, which often fail to capture the evolving and context-dependent nature of online mnemonics.

However, several challenges remain. Language in digital spaces evolves rapidly, with new slang, symbols, and coded expressions constantly emerging. While Content Moderation Model can identify many established patterns, it may miss newly coined mnemonics or context-specific meanings, especially when emojis or abbreviations are used in novel ways. The model's reliance on textual input also means it may fail to interpret images, memes, or blended multimedia content—modalities that are increasingly used to convey hidden or sensitive messages online. Furthermore, the nuanced and context-dependent use of mnemonics, particularly across different cultures and languages, continues to pose difficulties for even the most advanced language models.

9. References:

1. Stone, C.B. and Wang, Q. (2019), From Conversations to Digital Communication: The Mnemonic Consequences of Consuming and Producing Information via Social Media. *Top Cogn Sci*, 11: 774-793. <https://doi.org/10.1111/tops.12369>
2. Lee, Jaewook & Lan, Andrew. (2024). Exploring Automated Keyword Mnemonics Generation with Large Language Models via Overgenerate-and-Rank. 10.48550/arXiv.2409.13952.
3. Roediger, H. L. (1980). The effectiveness of four mnemonics in ordering recall. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5), 558–567. <https://doi.org/10.1037/0278-7393.6.5.558>
4. Gupta, Aditi & Kumaran, Ponpuraigam & Sureka, Ashish. (2012). Characterizing Pedophile Conversations on the Internet using Online Grooming.
5. Sesar, Kristina & Duda, Zita. (2019). Sexting and emotional regulation strategies among young adults. 10.6092/2282-1619.2019.7.2.18.
6. Hulstijn J. Mnemonic methods in foreign language vocabulary learning: Theoretical considerations and pedagogical implications. In: Coady J, Huckin T, eds. *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*. Cambridge Applied Linguistics. Cambridge University Press; 1996:203-224.
7. Adriansen M (2024). Latent and explicit mnemonic communities on social media: studying digital memory formation through hashtag co-occurrence analysis. *Memory, Mind & Media* 3, e13, 1–21. <https://doi.org/10.1017/mem.2024.7> Memory, Mind, and Media. <https://doi.org/10.1017/mem.2024.7> Published online by Cambridge University Press
8. Satadita Mookherjee, Smaraki Mohanty, Sourjo Mukherjee (2023). Affinity to mnemonic features of social media: Antecedent and Consequences. <https://doi.org/10.1002/cb.2210>
9. Ana Lúcia Migowski da Silva (2023), Mnemonic Practices on Social Media <https://doi.org/10.1007/978-3-658-41276-0>
10. Barber, S.J., Mather, M. Forgetting in context: The effects of age, emotion, and social factors on retrieval-induced forgetting. *Mem Cogn* 40, 874–888 (2012). <https://doi.org/10.3758/s13421-012-0202-8>
11. Obiora, Adanma & Uche, Anthony & Adikuru, Chinonso. (2025). Encoding and Decoding of Emoji on Social Media for Expression Sharing among University Students in Anambra State. *African Journal of Social Sciences and Humanities Research*. 8. 24-35. 10.52589/AJSSHR-IWUY1DYS.

12. S.Sumithra, Dr.P.Sujatha, "Evolution of Adult Content Detection" in the International Conference on Information and Business Intelligence (ICIBI – 2024) organized by Saveetha College of Liberal Arts & Sciences held on 15th Feb 2024
13. S.Sumithra, Dr.P.Sujatha, "Adult Content Detection with Machine Learning Techniques" in the International Conference On Tech Innovations with Multi-Disciplinary Approach (ICTIMDA - 2024) organized by Saveetha College of Liberal Arts & Sciences, Department of Computational Intelligence with Metharath University Thailand, ESN Publications & Computer Society of India, Chennai Chapter, held during 27 & 28th Feb 2024
14. S.Sumithra, Dr.P.Sujatha, "Performance Evaluation of Machine Learning Techniques and Large Language Models for Adult Content Detection" in Tamil Nadu State Council for Science and Technology on Artificial Intelligence and Education - Shaping the Future of Society (TNSCST - 2025) organized by Government Arts College, Nandanam, Chennai, India, held during 26th & 27th Feb 2025