DATA MINING
PROJECT PRESENTATION

IDENTIFICATION OF CRITICAL SUB-COMPONENTS IN CONVERSATIONS

GROUP 27 TEAM MEMBERS

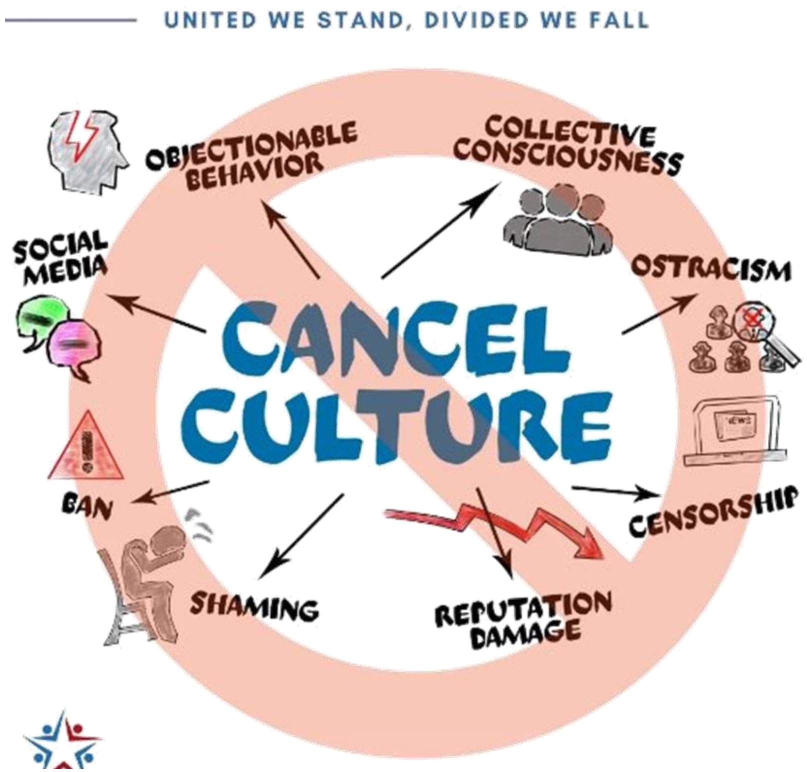
- SAGAR SINHA
- PRIYESHA JETHI
- SWATI KASHYAP
- DHRUVA MANISHKUMAR PATEL

AGENDA

- Problem Statement and Related Work
- Data Preprocessing
- Feature Extraction
- Annotation
- Data Modelling
- Key Results
- Limitations
- Proposed Solutions
- Future Work

PROBLEM STATEMENT AND RELATED WORK

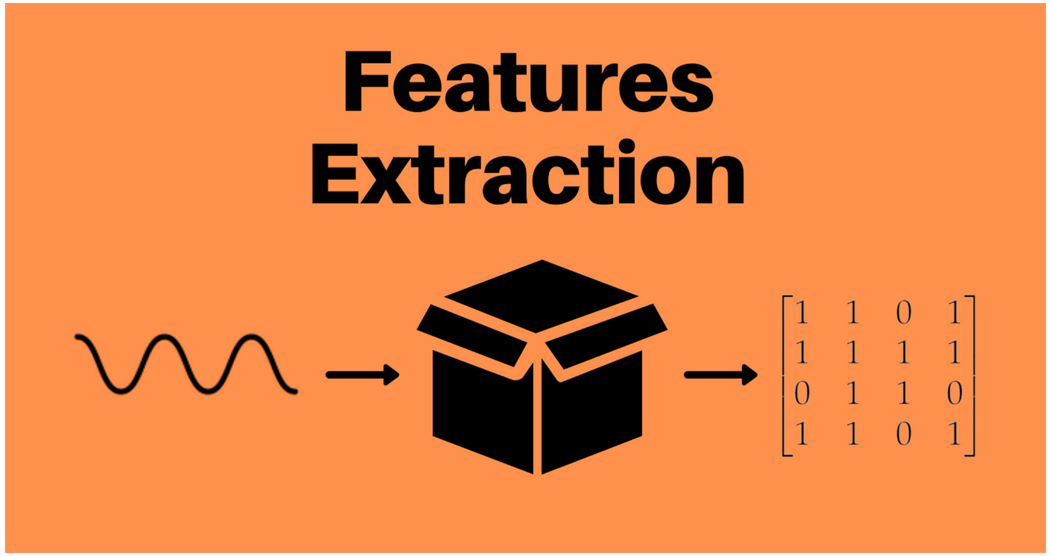
- The project aims to understand the complexities of Reddit conversations by analyzing both the language used and the conversational structure (using graph neural networks). This will help identify how these elements influence each other, leading to a comprehensive model of online discourse and strategies for promoting healthier online communities. Ultimately, addressing these challenges holds the potential to inform strategies for fostering healthier online communities and enhancing the quality of online interactions.
- Most of the existing studies have focused exclusively on individual comments/posts, ignoring the conversational context.



DATA PREPROCESSING

- Standardize data format by converting Convokit platform data into data frames for analysis
- Enhance data quality by filtering out bot comments and entries without speaker information
- Focus on higher-quality exchanges by excluding conversations with only a single level of depth
- Annotate preprocessed dataset(s) with appropriate labels
- Optimize data structure for compatibility with Graph Neural Networks (GNNs), facilitating efficient processing and analysis
- Implement a stratified data split approach for balanced training, validation, and testing datasets

FEATURE EXTRACTION



$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

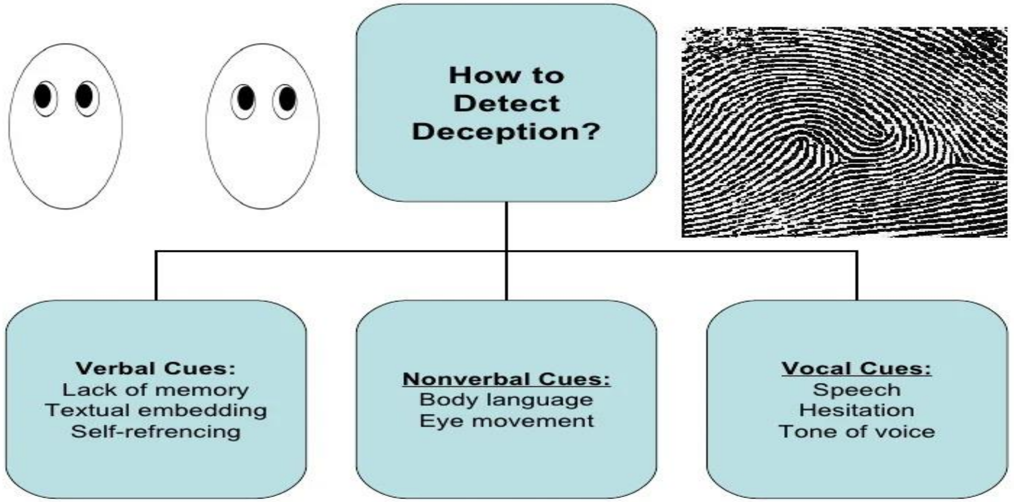
Hate Speech:

- Used sentiment analyzer by Vader for hate speech feature extraction. This tool effectively measures polarity in social media contexts
- Integrated Perspective API alongside the sentiment analyzer. Perspective API is designed to assess and address toxicity in online discourse
- These tools were chosen to improve the detection and comprehension of hate speech in online discussions

Trolls:

- Binary feature assignment: Comments categorized as either 1 or 0 based on specific criteria
  - Assigned a value of 1 if the comment meets certain conditions:
  - Contains child comments or direct replies with any instance of the word 'troll.'
  - Contains one or more words/phrases from a predefined list of offensive terms
- Comments meeting these criteria were marked as 1
- Comments not meeting these conditions were designated a binary feature value of 0

FEATURE EXTRACTION



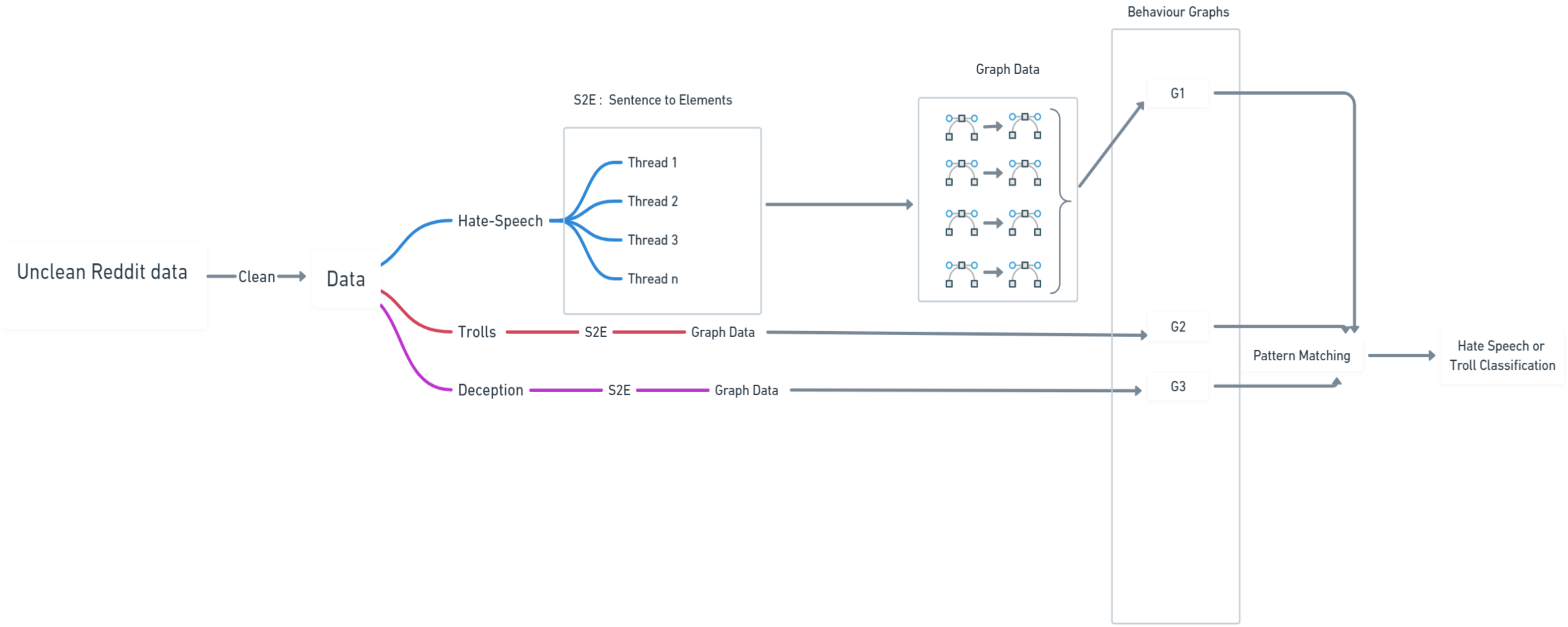
Deception:

- Convert text to lowercase, remove punctuations, stop words, links, emojis, and special characters, then tokenize and vectorize the text for feature extraction.
- Incorporate feature extraction steps for detecting deceptive clues within the corpus, including:
  - Counting modifiers (adjectives and adverbs)
  - Analyzing modal verbs
  - Tabulating self-references
  - Calculating subjectivity scores

ANNOTATION

- Due to the lack of annotated conversational data, we have performed an inhouse annotation scheme to detect node-edge relationships
- Three team members independently took turns to analyze sample conversations and provide individual annotations
- Fleiss Kappa, a statistical measure, was employed to assess the agreement among the annotators
- Till date, we have achieved an average score of 0.67 which we believe suggests substantial agreement among the annotators
- The process of inter-annotator agreement evaluation reinforces the reliability and consistency of the annotations obtained, indicating a strong foundation for the subsequent stages of analysis.

DATA MODELLING



KEY RESULTS

GNN Model Accuracy/ GCN Graph Convolutional Network

Model	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
Hate Speech	0.90	0.1	0.78	0.45
Trolls	0.87	0.2	0.72	0.48

LIMITATIONS IN WORK

- Analyzing comments in isolation, without their broader conversational context, limits understanding and risks misinterpretation of their meaning.
- Identifying trolling behavior is difficult because it depends heavily on context, varying across online communities and often lacking the obvios signals present in hate speech.
- Detecting deception on Reddit is complex because this subproblem goes beyond blatant lies, encompassing subtle omissions and manipulations, and Reddit platform's features (anonymity, diverse language use, etc.) futher complicate identification

PROPOSED SOLUTIONS

- Keeping the entire conversation together in the annotation process could provide crucial context for accurate labeling and a deeper understanding of conversational dynamics.
- Understanding trolling versus hate speech nuances can possibly inform more effective moderation strategies and content policies.
- Employ a combined approach of linguistic analysis, conversational context evaluation, user history analysis, and subreddit-specific tailoring for nuanced deception detection on Reddit.

FUTURE WORK

- Employing advanced GNN architectures allows for deeper analysis of hierarchical structures, temporal shifts, and multimodal elements, ultimately enhancing the detection of harmful online behaviors.
- Including longer, in-depth conversations with edge-level annotations reveals intricate interaction patterns, providing more robust training data for a GNN to accurately detect subtle nuances of hate speech, trolling, and deception
- Including edge-level annotation can provide explicit information about the nature of the relationship between comments (e.g., agreement, disagreement, clarification, question-answer). This allows the GNN to learn more nuanced interaction patterns, going beyond just the content of individual comments