# Analyzing News Sentiment on YouTube and Snapchat's Impact on Youth Mental Health

**Joseph Keepers, Kayla DePalma**
jkeepers@stevens.edu | kdepalma@stevens.edu

## 1 Introduction

This project investigates how news media portray the relationship between video-based social media platforms, particularly YouTube and Snapchat, and youth mental health. The primary objective is to conduct stance detection and sentiment analysis to determine whether articles frame the issue positively (supportive, destigmatizing, hopeful), negatively (stigmatizing, fear-based, alarming), neutrally, or in a mixed way where both positive and negative perspectives coexist [1]. A dataset of news articles is collected using keyword searches related to youth mental health and video-based social media. Because keyword-based retrieval often captures irrelevant results, preprocessing is conducted through large language model (LLM)-assisted filtering to ensure that only relevant articles are retained. All articles are evaluated by Gemini, GPT, and Grok, which classify each as relevant or borderline to the topic. The results are then manually reviewed to finalize the set of relevant articles. The relevant articles are then annotated for sentiment using a human-machine collaboration approach, in which LLM-generated predictions are validated and refined by human reviewers.

The workflow proceeds with data preparation, including preprocessing, dataset splitting, and descriptive statistics, followed by text-to-vector representation through vocabulary construction, co-occurrence matrices, and dimensionality reduction for visualization. In addition to annotation, multiple LLMs (e.g., ChatGPT, Gemini, Grok) are compared to evaluate their effectiveness in filtering and labeling. Statistical analyses are conducted to compare sentiment distributions across platforms and classes. Finally, classification models, including Logistic Regression, LSTM, and fine-tuned BERT, are trained and assessed to identify sentiment patterns, with model performance compared across approaches. The anticipated outcomes include determining whether negative narratives dominate, positive and supportive framings are emerging, neutral coverage remains prevalent, or mixed perspectives are increasingly common. Ultimately, the project provides a systematic view of how youth mental health in relation to digital platforms is represented in news discourse and offers insights into broader challenges in automated stance detection and media analysis.

A key challenge encountered during dataset construction was the substantial reduction in data after filtering. Many articles retrieved through keyword searches were duplicates, in a different language, or unrelated to the target topic. As a result, a significant portion of the initial YouTube and Snapchat datasets were removed during preprocessing and LLM-assisted relevance evaluation, leaving a much smaller set of high-quality articles. While this filtering improved the precision of the datasets, it also limited the sample size available for later analysis and model development. To address this limitation, we combined the YouTube and Snapchat datasets, increasing the number of articles available for training and ensuring sufficient coverage for sentiment classification.

Previous work has demonstrated the usefulness of sentiment analysis in tracking the polarity of news articles about topics such as politics, economics, and public health [2, 3]. Such studies show that sentiment analysis of news is valuable because media outlets play a significant role in shaping public opinion [4]. However, a key challenge in sentiment analysis is granularity. Models may classify sentiment at the document, sentence, or aspect level, and each approach has limitations. Context is especially important in longer texts, like news articles, where multiple entities may be discussed with differing sentiments, requiring careful consideration [5]. In the context of mental

health, most sentiment analysis research has focused on user-generated content from platforms like Twitter [6], which provides insight into individuals' personal experiences and expressions. This project expands on that line of work by shifting attention from user content to media narratives, examining how news coverage itself portrays social media platforms in relation to youth mental health. To address the challenge of fine-grained sentiment detection in long, context-rich texts, we classify articles into four sentiment categories: positive, negative, neutral, and mixed, and evaluate the task using multiple existing models, including logistic regression, LSTM, and BERT. Our contributions include benchmarking these models on this novel news-based dataset, assessing the impact of different modeling techniques, and identifying strategies that yield optimal performance for sentiment classification in this domain.

## 2 Problem Formulation

### 2.1 Notation

Let each news article be represented as a token sequence

$$x = (a_1, a_2, \ldots, a_n) \in X.$$

Define the set of stance labels

$$Y = \{y_{\text{pos}},\, y_{\text{neg}},\, y_{\text{neu}},\, y_{\text{mix}}\},$$

where $y_{\text{mix}}$ denotes mixed framing (both positive and negative perspectives). Let the annotated dataset be

$$D = \{(x_i, y_i^*)\}_{i=1}^{N},$$

where $y_i^*$ is the final human-refined label after LLM prelabeling and human correction.

### 2.2 Task A: LLM Comparison

After using multiple LLMs (ChatGPT, Gemini, Grok, etc.) to generate preliminary labels for the dataset, their outputs will be compared both against one another and against the final human-refined labels. This comparison will be presented using confusion matrices and performance tables to highlight agreement, disagreement, and overall alignment with the ground truth, defined here as the human-refined labels.

**Example Confusion Matrix**

| True \ Pred | Pos | Neg | Neu | Mix |
|---|---|---|---|---|
| Pos | 12 | 3 | 2 | 1 |
| Neg | 4 | 18 | 1 | 0 |
| Neu | 2 | 1 | 15 | 2 |
| Mix | 1 | 0 | 3 | 10 |

Table 1: Example confusion matrix for an LLM.

**Example Performance Table**

| Model | Accuracy | Macro Precision | Macro Recall | Macro F1 |
|---|---|---|---|---|
| ChatGPT | 0.82 | 0.80 | 0.79 | 0.79 |
| Gemini | 0.78 | 0.76 | 0.74 | 0.75 |
| Grok | 0.74 | 0.72 | 0.71 | 0.71 |

Table 2: LLM Agreement with Human-Refined Sentiment Labels.

### 2.3 Task B: Statistical Comparison of Classes

We report descriptive statistics for stance distribution across the dataset and optionally by platform.

**Example Class Distribution Table**

| Class | Count | Percentage | Avg. Article Length (tokens) |
|---|---|---|---|
| Pos | 120 | 24% | 580 |
| Neg | 210 | 42% | 610 |
| Neu | 100 | 20% | 540 |
| Mix | 70 | 14% | 600 |

Table 3: Example class distribution with descriptive statistics.

### 2.4 Task C: Supervised Multi-class Classification

Learn a classifier

$$f_\theta : X \to \Delta(Y)$$

that outputs a probability distribution over the four classes. Use the cross-entropy loss:

$$\mathcal{L}(\theta) \;=\; -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in Y}\mathbf{1}(y_i^* = c)\,\log p_\theta(c \mid x_i),$$

and obtain parameters by

$$\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta).$$

Evaluate $f_{\hat{\theta}}$ on a held-out test set using confusion matrices and standard metrics (accuracy, per-class precision/recall/F1, macro F1).

## 3 Methods

### 3.1 Data Collection and Filtering

The datasets were given in the pre-defined project for Stevens Institute of Techonology course CS 584 Natural Language Processing taught by Professor Ping Wang. News articles were collected via keyword searches related to YouTube, Snapchat, and youth mental health. For reference, see the following sources:

- YouTube Data
- Snapchat Data

To ensure relevance, we employ **large language models (LLMs)** to screen articles. Duplicate articles, non-English articles, and articles not related to the topic will be removed:

- Automate the removal of duplicate and non-English articles to improve dataset quality and prevent clearly irrelevant items from being passed to the LLMs
- Task-specific prompts check for mentions of both video-based social media and youth mental health.
- LLMs flag borderline cases for human verification, ensuring a clean dataset.

**Innovative aspect:** Combining automated LLM filtering with human validation accelerates preprocessing while maintaining high-quality relevance.

**Duplicate and Non-English Article Removal**

Removing duplicates proved more challenging than initially expected because many articles were not exact copies (e.g. differing by a single word or sentence). To address this, a two-step approach was used.

First, we applied `pandas`' `drop_duplicates` function to remove articles with identical title and body text, keeping only the first occurrence. Since this did not capture articles with slightly varied content, we implemented a second step: we extracted the first five words from each article title and

applied `drop_duplicates` again, this time based solely on these first five words. Manual inspection of the YouTube and Snapchat datasets confirmed that this approach effectively removed near-duplicate articles, where titles differed slightly due to the news outlet but the body text was essentially the same.

In addition to duplicates, non-English articles were removed to ensure language consistency for model development. Using `langdetect`'s `detect` function, each article title was assigned a language identifier, and only English-language articles were retained.

After filtering, 604 of 1,326 YouTube articles remained (705 duplicates and 17 non-English articles removed), and 250 of 676 Snapchat articles remained (424 duplicates and 2 non-English articles removed).

### LLM Relevancy Filtering

We used three LLMs (`gemini-2.0-flash`, `gpt-4o-mini`, and `grok-4`) to classify articles as either relevant or borderline. These models were selected because prior benchmarking and comparison studies demonstrate meaningful differences in their performance, with each model showing strengths in slightly different applications [7, 8]. By leveraging all three models and comparing their outputs, we aimed to obtain a more robust and representative classification of article relevance and later for sentiment classification.

Each article from the YouTube and Snapchat datasets was passed to the LLMs with the following prompt:

```
"""
You are a research assistant. Determine if the following article is relevant to:
"video-based social media (YouTube, Snapchat) and youth mental health."

Respond in JSON with:
- "relevant": true/false
- "borderline": true/false
- "reason": short explanation

Article:
{excerpt}
"""
```

This prompt was designed to balance simplicity with specificity. By framing the LLM as a research assistant, it provided context about our goals and encouraged careful consideration rather than arbitrary decisions. The JSON output structure allowed the LLMs to classify each article while also providing a concise explanation for their decision, which allows for efficient manual review and ensures transparency in the classification process. Identifying articles as borderline is important because it flags content that may be relevant but requires human review to confirm, ensuring that potentially useful articles are not mistakenly excluded.

The LLM outputs revealed patterns in relevance classification. For YouTube articles, all three models unanimously agreed on 238 relevant articles out of 604. In addition, Grok identified 48 more relevant articles, Gemini 91 more, and GPT 73 more. For Snapchat articles, unanimous agreement was on 231 articles out of 250, with Grok identifying 6 additional relevant articles, Gemini 14, and GPT 5. Many of the articles additionally marked as relevant were recognized by two of the three LLMs, but not by all three, highlighting partial agreement among the models. To ensure accuracy, we manually validated the articles marked as relevant by all three LLMs and confirmed their true relevance. We also reviewed the other articles labeled as relevant or borderline by one or two models, selecting those we deemed relevant to add back into the dataset.

Analysis of these results shows that Gemini tended to classify the most articles as relevant, while Grok was the most conservative. After manual review, it appeared that Gemini and GPT often agreed on which articles were relevant, whereas Grok frequently disagreed. In particular, articles identified as relevant by Grok were less likely to be recognized as relevant by Gemini or GPT. This highlights differences in LLM sensitivity and suggests that combining multiple LLM outputs can provide a more balanced and thorough filtering process.

The LLMs identified 238 of 604 YouTube articles and 231 of 250 Snapchat articles as relevant. This suggests that Snapchat articles were generally more focused on youth mental health and video-based social media, whereas a larger proportion of YouTube articles were less relevant or borderline and required filtering. One reason for this is that some articles were included in the initial YouTube dataset simply because they mentioned "YouTube" and discussed mental health, even if they did not emphasize YouTube's impact on youth mental health, so they were excluded from our relevant dataset.

Both datasets experienced a significant reduction in articles following preprocessing and relevance filtering. To preserve a dataset large enough for effective sentiment analysis, we combined the YouTube and Snapchat articles, increasing the total number of samples available for training and evaluation. After merging the datasets and removing duplicate entries, the final dataset contains 401 articles.

### 3.2 Annotation and Labeling

Each article is assigned a sentiment/stance label: Positive, Negative, Neutral, or Mixed.

- **Human–Machine Collaboration:** LLMs first generate predicted labels (zero-shot), which human reviewers validate and refine.
- **Tools:** ChatGPT, Gemini, Grok for initial labeling; spreadsheets or annotation tools for human review.

**Innovation:** Evaluating multiple LLMs on the same annotation task to compare their effectiveness.

We used the same three LLMs as before (`gemini-2.0-flash`, `gpt-4o-mini`, and `grok-4`) to label each article's sentiment as positive, negative, neutral, or mixed, based on how the article portrayed the impact of video-based social media on youth mental health.

Each article from the dataset of relevant articles was passed to the LLMs with the following prompt:

```
"""
You are an experienced data analyst tasked with evaluating the sentiment of
news articles discussing the impact of video-based social media on youth
mental health. The article may not explicitly state its sentiment, so you must
infer it based on tone, framing, and language.

Your task is to classify the overall sentiment of each article into one of
four categories: POSITIVE, NEGATIVE, MIXED, or NEUTRAL.

Use the following criteria:
- POSITIVE: At least 75% of the language expresses optimism, support,
            destigmatization, or hopefulness.
- NEGATIVE: At least 75% of the language conveys fear, stigma, alarm, or a
            harmful portrayal.
- MIXED: The article presents both positive and negative perspectives in
         roughly equal balance (around 50/50).
- NEUTRAL: The article maintains an objective or factual tone, with no clear
           positive or negative stance.

Respond in JSON with:
- "sentiment": "POSITIVE" | "NEGATIVE" | "MIXED" | "NEUTRAL",
- "reason": "Brief justification (1-2 sentences)"

Article:
{excerpt}
"""
```

The prompt was designed to provide clear task instructions and consistent sentiment definitions, reducing ambiguity across models. By outlining specific criteria for each sentiment category and requiring a short justification, it ensures that classifications are both interpretable and grounded

in textual evidence. The JSON format also allows for structured and reproducible outputs. The instruction to infer tone and framing ensures deeper contextual analysis rather than surface-level keyword matching.

After gathering the LLM outputs, we compared their classifications and applied human validation to create the final annotated dataset.

### 3.3 LLM Comparison (Task A)

After multiple LLMs (ChatGPT, Gemini, Grok) are prompted to generate preliminary sentiment labels for the dataset, their predictions are compared both against each other and against the human-refined ground truth. Results are summarized using confusion matrices and performance tables, with emphasis on identifying the model whose labels align most closely with the human-refined labels.

- **Tools:** Python, `scikit-learn` metrics, `pandas` for table generation.

### 3.4 Data Preparation

- **Preprocessing:** Text cleaning (lowercasing, removing punctuation, URLs, and numbers, expanding contractions, removing possessive "'s", and normalizing whitespace), followed by tokenization, and optional or selective stopword removal. Sentiment-changing stopwords such as "not," "no," and "never" are preserved, as they directly influence sentiment polarity.
- **Data Splitting:** Stratified split into 80% training, 10% validation, and 10% test sets.
- **Statistics:** Article count per class.
- **Tools:** Python libraries such as `NLTK`, `numpy`, `sklearn`, and `pandas`.

### 3.5 Statistical Analysis of Classes (Task B)

We examine the distribution of stance classes across the dataset and across platforms (YouTube vs. Snapchat).

- Frequency tables for class counts and percentages.
- Descriptive statistics, e.g., mean article length per class, proportion comparisons.
- **Tools:** `pandas` for summaries, `matplotlib`/`seaborn` for visualizations.

### 3.6 Text Representation

- TF-IDF vectors used with logistic regression.
- Word embeddings (task-specific embeddings learned directly from LSTM's embedding layer and BERT transformer embeddings).
- Co-occurrence-based semantic representations, using both the standard co-occurrence matrix and an enhanced version computed with Positive Pointwise Mutual Information (PPMI).
- **Dimensionality Reduction:** SVD to visualize word clusters.
- **Visualization Tools:** `matplotlib`, `seaborn`, `TSNE`.

### 3.7 Classification Models (Task C)

- **Logistic Regression (Baseline, Classical ML):** Simple, interpretable, works well with TF-IDF features. Fast training, coefficients indicate important words. Limited context understanding.
- **LSTM (Neural Network, Sequential Modeling):** Captures sequential dependencies, handles longer texts, commonly benchmarked. Slower to train, less effective than transformers on small datasets.
- **Fine-tuned BERT (Transformer, State of the Art):** Pretrained on large corpora, captures context and nuances. High performance in modern sentiment tasks. Requires more compute and GPU for training.

- **Evaluation Metrics:** Accuracy, precision, recall, F1 (per class and macro F1).
- **Tools:** `scikit-learn`, `PyTorch`, `TensorFlow`, `Hugging Face Transformers`.

Previous studies have compared logistic regression, LSTM, and BERT for sentiment analysis, showing that logistic regression is a strong classical baseline, LSTMs offer advantages for sequence modeling, and BERT, with its pretrained bidirectional transformer architecture, achieves superior performance on text classification tasks [9]. We included BERT as a model because transformer-based architectures have been shown to capture nuanced sentiment more effectively than traditional machine learning approaches [10]. Building on these findings, we selected these three models to compare their performance on news articles and evaluate their effectiveness for sentiment analysis in the context of youth mental health.

However, prior work has also shown that BERT's performance can degrade when fine-tuned on small or imbalanced datasets. In particular, transformer-based models tend to overfit quickly when the number of training examples is limited, and their ability to distinguish minority sentiment classes suffers as a result. Studies demonstrate that expanding the size of the fine-tuning dataset substantially improves classifier sensitivity and F1-scores, especially for underrepresented classes [11]. Given the relatively small size of our dataset, these limitations pose a challenge. Nonetheless, the literature suggests that with a larger or more balanced training set, BERT would likely achieve stronger and more stable performance.

### 3.8 Innovative Approaches

- **Human–LLM Collaboration:** Efficiently combines automated labeling with human review for high-quality annotations.
- **Cross-LLM Evaluation:** Comparing multiple LLMs provides insights into their consistency and reliability as annotation aids for sentiment labeling.
- **Word Relationship Visualization:** Co-occurrence matrices and dimensionality reduction give a richer understanding of semantic patterns in media framing.

## 4 Dataset and Experiments

### 4.1 Datasets

The primary dataset consists of news articles related to video-based social media (YouTube and Snapchat) and youth mental health.

- **Sources:** Online news outlets, collected via keyword searches.
- **Annotations:** Each article is labeled with one of four stance classes: Positive, Negative, Neutral, or Mixed. Labels are refined through human–LLM collaboration.
- **Data Split:** Stratified into 80% training, 10% validation, and 10% test sets.

The finalized dataset contains 401 articles relating to both YouTube and Snapchat. Table 11 summarizes the class label distribution and reports descriptive statistics, such as the average article length in tokens. The vocabulary consisted of the top occuring 5000 words.

Before training our models, we applied several preprocessing steps to the article text. The body of each article was converted to lowercase, contractions were expanded, and punctuation, numbers, URLs, excess whitespace, and possessive "'s" were removed. Because the task involves sentiment analysis, it was essential to preserve sentiment-bearing terms such as "no," "never," and "not," which can reverse or significantly alter the meaning of a sentence. Expanding contractions ensured that negations (e.g., don't $\rightarrow$ do not) were explicitly represented. Stopwords were removed only for the logistic regression model, excluding sentiment-bearing terms, because this preprocessing improved its predictive accuracy. For the LSTM and BERT models, stopwords were retained to preserve contextual information. Finally, we removed unnecessary punctuation and normalized whitespace to eliminate artifacts from the data collection process, such as extraneous newline markers, which would otherwise introduce noise into the model.

For each model, we applied stratified sampling to split the dataset into 80% training, 10% validation, and 10% testing. Because the class distribution was imbalanced, stratification ensured that each split

preserved similar proportions of the four sentiment classes, making the training set representative of the validation and test sets. This was important for obtaining reliable performance estimates and for helping the models learn patterns that generalize. The validation set was used for hyperparameter tuning, as its performance closely reflects how the models would perform on the unseen test set.

### 4.2 Experiments

**Task A: LLM Comparison**

- **LLMs:** ChatGPT, Gemini, Grok.
- **Experiments:** Preliminary sentiment label predictions on the unlabeled dataset, compared to one another and to the human-refined ground truth labels.
- **Evaluation Metrics:** Confusion matrices, accuracy, per-class precision, recall, F1 score; highlight the LLM most consistent with human-refined labels.

The three models unanimously agreed on the sentiment classification of 289 out of 401 articles, which is roughly 72% agreement. This demonstrates strong baseline consistency among the models and suggests that the majority of articles have sentiment that is clear and unambiguous.

As shown in Tables 4–6, all three models rarely misclassified positive articles as negative or neutral, demonstrating strong recognition of positive sentiment. Most negative articles were correctly identified, though some were misclassified as mixed, suggesting that certain articles labeled purely negative by humans contain nuanced tones (e.g., Table 4 shows GPT misclassifying 38 negative articles as mixed). Neutral sentiment proved the most challenging for all models. For instance, GPT misclassified 11 neutral articles as negative and 15 as mixed (Table 4), highlighting the difficulty in detecting purely neutral language. Mixed sentiment was generally well-identified, particularly by GPT and Gemini, though occasional errors occurred where mixed articles were predicted as negative or neutral (Tables 4 and 6), reflecting the subtlety and complexity of language that LLMs sometimes oversimplify.

**Confusion Matrices for LLMs vs Human-Refined Sentiment Labels**

| True \ Pred | Pos | Neg | Neu | Mix |
|---|---|---|---|---|
| Pos | 54 | 0 | 0 | 3 |
| Neg | 3 | 216 | 1 | 38 |
| Neu | 4 | 11 | 13 | 15 |
| Mix | 0 | 0 | 0 | 43 |

Table 4: Confusion Matrix for GPT vs Human-Refined Labels.

| True \ Pred | Pos | Neg | Neu | Mix |
|---|---|---|---|---|
| Pos | 53 | 0 | 1 | 3 |
| Neg | 5 | 199 | 25 | 28 |
| Neu | 2 | 1 | 38 | 2 |
| Mix | 2 | 1 | 4 | 36 |

Table 5: Confusion Matrix for Grok vs Human-Refined Labels.

| True \ Pred | Pos | Neg | Neu | Mix |
|---|---|---|---|---|
| Pos | 54 | 1 | 1 | 1 |
| Neg | 9 | 238 | 1 | 9 |
| Neu | 4 | 9 | 20 | 10 |
| Mix | 0 | 4 | 2 | 37 |

Table 6: Confusion Matrix for Gemini vs Human-Refined Labels.

**Performance Table**

| Model | Accuracy | Macro Precision | Macro Recall | Macro F1 |
|---|---|---|---|---|
| GPT | 0.813 | 0.800 | 0.772 | 0.717 |
| Grok | 0.815 | 0.731 | 0.856 | 0.772 |
| Gemini | 0.873 | 0.808 | 0.800 | 0.786 |

Table 7: LLM Agreement with Human-Refined Sentiment Labels.

As shown in Table 7, Gemini achieves the highest overall accuracy (0.873) and the highest Macro F1 (0.786), with strong Macro Precision (0.808) and Recall (0.800), indicating it is the most balanced model for overall sentiment classification. Grok has slightly lower accuracy (0.815) but the highest Macro Recall (0.856), meaning it is particularly good at identifying true positives across all sentiment categories, including less frequent or subtle labels. ChatGPT has the lowest overall performance, with an accuracy of 0.813 and Macro F1 of 0.717, suggesting it struggles to balance precision and recall across classes, especially for nuanced categories such as neutral and mixed sentiment.

These results highlight complementary strengths among the models. Gemini's strong overall accuracy and balanced F1 indicate it is reliable for general classification, while Grok's high recall makes it sensitive to detecting underrepresented sentiments. ChatGPT's weaker F1 underscores its difficulty in capturing subtle sentiment distinctions. Overall, combining insights from these models could help mitigate individual weaknesses and improve coverage of nuanced sentiment in the dataset.

**Task B: Statistical Analysis of Classes**

- **Experiments:** Compare stance distributions, including class counts, percentages, and mean article lengths.
- **Metrics:** Frequency tables, descriptive statistics, simple proportion comparisons to highlight differences across platforms.

**Overall Class Distribution Table**

| Class | Count | Percentage | Avg. Article Length (tokens) |
|---|---|---|---|
| Pos | 61 | 15.21% | 433 |
| Neg | 227 | 56.61% | 523 |
| Neu | 14 | 3.49% | 475 |
| Mix | 99 | 24.69% | 666 |

Table 8: Class distribution with descriptive statistics for GPT.

| Class | Count | Percentage | Avg. Article Length (tokens) |
|---|---|---|---|
| Pos | 62 | 15.46% | 444 |
| Neg | 201 | 50.12% | 567 |
| Neu | 68 | 16.96% | 517 |
| Mix | 69 | 17.21% | 584 |

Table 9: Class distribution with descriptive statistics for Grok.

| Class | Count | Percentage | Avg. Article Length (tokens) |
|---|---|---|---|
| Pos | 67 | 16.71% | 442 |
| Neg | 252 | 62.84% | 544 |
| Neu | 24 | 5.99% | 456 |
| Mix | 57 | 14.21% | 678 |

Table 10: Class distribution with descriptive statistics for Gemini.

| Class | Count | Percentage | Avg. Article Length (tokens) |
|---|---|---|---|
| Pos | 57 | 14.21% | 449 |
| Neg | 258 | 64.34% | 553 |
| Neu | 43 | 10.72% | 533 |
| Mix | 43 | 10.72% | 615 |

Table 11: Class distribution with descriptive statistics for Human-Refined Sentiment Labels.

The class distribution across the dataset reveals a consistent pattern of sentiment prevalence. Tables 8-11 demonstrate that negative sentiment is the most frequent across all models and human-refined labels, ranging from approximately 50% (Grok) to over 64% (human labels), reflecting the prevalence of articles highlighting concerns or risks associated with video-based social media. Positive sentiment is relatively underrepresented, comprising roughly 14–17% of articles across all models. Neutral articles are particularly rare in GPT and Gemini outputs (3–6%), though Grok and the human-refined labels show a slightly higher proportion (17% and 11%, respectively), suggesting some variation in how neutral content is interpreted. Mixed sentiment is moderately represented, ranging from 10% to 25%, with GPT identifying the highest proportion (24.7%).

Average article lengths vary by sentiment and model, with mixed sentiment articles generally longer (ranging from 578 to 678 tokens), potentially reflecting the complexity and nuance in articles containing multiple perspectives. Negative articles tend to be slightly longer than positive or neutral articles, which may indicate more in-depth discussion of risks or concerns. These descriptive statistics emphasize class imbalance, highlighting that negative sentiment dominates the corpus and that careful attention is needed to evaluate model performance on underrepresented classes such as neutral and positive articles.

**Task C: Classification Models**

- **Models:** Logistic Regression, LSTM, Fine-tuned BERT
- **Parameters:** Regularization (LR), learning rate, hidden layers, dropout (NNs), embedding type, fine-tuning strategy (BERT)
- **Evaluation Metrics:** Accuracy, per-class precision, recall, F1, macro F1, confusion matrices

**Semantic Analysis**

After preprocessing the text, removing non-sentiment-bearing terms, and tokenizing, we constructed co-occurrence matrices using raw counts as well as Positive Pointwise Mutual Information (PPMI) to highlight meaningful associations rather than just frequent words. We explored these matrices using visualization techniques such as heatmaps and word clustering to better understand relationships between words in the news articles.

As expected, the most frequent words and clusters reflected the domain of the dataset: names of social media platforms (e.g., YouTube, Snapchat, TikTok) clustered together, and terms like young, social media, and mental health were highly frequent. Heatmaps showed predictable co-occurrences, such as social with media and platform names co-occurring with each other.

Dimensionality reduction using SVD on the PPMI matrix produced similarly predictable associations, though it offered limited additional insight, likely due to the relatively small corpus of 401 articles and the limited vocabulary diversity. Interestingly, words like not and like appeared together in a cluster, reflecting the prevalence of negative sentiment in the dataset.

To further explore semantic patterns, we applied t-SNE to the 50 most frequent words from the PPMI matrix. The resulting plot also showed expected clusters: social media platforms grouped together, words like district, school, and boards formed a cluster, content and users clustered, and kids, teens, parents, and online appeared together.

These results are consistent with expectations because PPMI emphasizes the most frequent co-occurrences in the corpus. In our case, the dominance of terms related to social media and youth mental health tends to overshadow subtler patterns. Nevertheless, the analysis confirms that the articles are highly relevant to our topic and capture the main semantic relationships in the text.

**Logistic Regression**

We trained logistic regression models both with and without stratified sampling to evaluate the effect of stratification on model performance and to get a baseline for our sentiment analysis task.

After preprocessing the text data, stopwords were removed (except for sentiment-bearing terms such as "not," "no," and "never"). The cleaned text was transformed into TF-IDF vectors using a maximum of 20,000 features and n-grams of size 1 and 2. The articles were tokenized using the TF-IDF vectorizer. We performed hyperparameter tuning on the regularization parameter $C$, testing values of 0.01, 0.1, 1, 10, and 100. Each model was trained on the training set, evaluated on the validation set, and the $C$ value yielding the highest validation accuracy was selected. The final model was then trained using this optimal $C$ and evaluated on the test set. L2 regularization was applied with 'solver='lbfgs'' and 'max_iter=1000'.

For logistic regression **without stratified sampling**, the best $C$ was 100, achieving a validation accuracy of 0.875. The test set evaluation metrics are reported in Table 11.

For logistic regression **with stratified sampling**, the optimal $C$ was 1, with a validation accuracy of 0.775. Test set metrics are shown in Table 12.

| Class | Precision | Recall | F1-score | support |
|---|---|---|---|---|
| Mix | 0.67 | 0.40 | 0.50 | 5 |
| Neg | 0.71 | 1.00 | 0.83 | 22 |
| Neu | 1.00 | 0.20 | 0.33 | 5 |
| Pos | 0.83 | 0.56 | 0.67 | 9 |
| Accuracy | - | - | 0.73 | 41 |
| Macro avg | 0.80 | 0.54 | 0.58 | 41 |
| Weighted avg | 0.77 | 0.73 | 0.69 | 41 |

Table 12: Classification report for logistic regression without stratified sampling.

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Mix | 0.00 | 0.00 | 0.00 | 5 |
| Neg | 0.70 | 1.00 | 0.83 | 26 |
| Neu | 1.00 | 0.25 | 0.40 | 4 |
| Pos | 1.00 | 0.50 | 0.67 | 6 |
| Accuracy | - | - | 0.73 | 41 |
| Macro avg | 0.68 | 0.44 | 0.47 | 41 |
| Weighted avg | 0.69 | 0.73 | 0.66 | 41 |

Table 13: Classification report for logistic regression with stratified sampling.

Without stratified sampling, the model was able to correctly label at least one article from all sentiment classes, whereas with stratification, it failed to identify mixed sentiment articles. However, the stratified model consistently identified neutral and positive articles correctly, whereas the non-stratified model only fully identified neutral articles. Despite these differences, both models achieved an overall accuracy of 73%.

Although the non-stratified model shows slightly higher metrics, we report results using **stratified sampling** as the baseline. Stratification ensures that each class is proportionally represented in the training, validation, and test sets, preventing minority classes from being underrepresented, which could artificially inflate performance metrics. This approach provides a more realistic and reliable evaluation, particularly for less frequent sentiment classes.

Tables 14 and 15 summarize the class distributions across the split datasets. They confirm that stratified sampling produces a more representative split, whereas the non-stratified split shows more variation in the number of negative and positive samples, which can impact model performance.

| Class | Training | Validation | Testing |
|---|---|---|---|
| Mix | 34 | 4 | 5 |
| Neg | 209 | 27 | 22 |
| Neu | 34 | 4 | 5 |
| Pos | 43 | 5 | 9 |

Table 14: Class distribution across split datasets for non-stratified sampling.

| Class | Training | Validation | Testing |
|---|---|---|---|
| Mix | 34 | 4 | 5 |
| Neg | 206 | 26 | 26 |
| Neu | 34 | 5 | 4 |
| Pos | 46 | 5 | 6 |

Table 15: Class distribution across split datasets for stratified sampling.

**LSTM**

To evaluate the effect of bidirectionality on model performance, we trained both a standard LSTM and a Bidirectional LSTM (BiLSTM) on the dataset. We wanted to investigate which model would perform better for our sentiment analysis task. The data was split into training, validation, and test sets using stratified sampling to ensure that each sentiment class was proportionally represented. Since LSTMs require integer labels, we used `sklearn`'s `LabelEncoder` to map the four sentiment classes to integers. The text was then tokenized and padded to create sequences of uniform length. Both models used task-specific embeddings learned directly from the embedding layer.

**LSTM Architecture:** The LSTM model consisted of two stacked LSTM layers with 64 and 32 units, respectively, followed by a dropout layer with a rate of 0.5 and a dense output layer with 4 units and a softmax activation:

```
LSTM_model = Sequential([
    Embedding(vocab_size, 128, mask_zero=True, trainable=True),
    LSTM(64, return_sequences=True),
    LSTM(32, return_sequences=False),
    Dropout(0.5),
    Dense(4, activation='softmax')
])
```

The model was trained using the Adam optimizer and `sparse_categorical_crossentropy` loss for 15 epochs with a batch size of 32. Hyperparameter tuning included adjusting the number of LSTM layers, dropout layers and rates, and the number of epochs. The configuration above yielded the best validation performance. The test set evaluation metrics are reported in Table 16.

**Bidirectional LSTM Architecture:** The BiLSTM model used a single bidirectional LSTM layer with 64 units, followed by dropout, a dense ReLU layer with 32 units, another dropout layer, and a dense output layer with softmax activation:

```
BiLSTM_model = Sequential([
    Embedding(vocab_size, 128, mask_zero=True, trainable=True),
    Bidirectional(LSTM(64, return_sequences=False)),
    Dropout(0.5),
    Dense(32, activation='relu'),
    Dropout(0.5),
    Dense(4, activation='softmax')
])
```

It was trained using Adam and `sparse_categorical_crossentropy` for 13 epochs with a batch size of 32. Similar hyperparameter tuning was performed, and this configuration achieved the best results on the validation set. The test set evaluation metrics are shown in Table 17.

**Performance Analysis:** The BiLSTM model outperformed the standard LSTM in overall metrics. It achieved higher accuracy (0.80 vs 0.76), macro and weighted averages for precision, recall, and F1-score, suggesting that incorporating bidirectionality helps the model better capture context from both past and future tokens in the text.

Looking at class-level performance:

- For both models, the negative sentiment class was predicted with high recall, likely because it dominates the dataset.
- The LSTM struggled to identify mixed and neutral sentiment articles, particularly under-predicting the mixed class.
- The BiLSTM improved performance on neutral and positive articles, with more balanced F1-scores across classes, although the mixed class remains challenging due to the limited number of samples.

Both models achieved a similar AUC score of 0.84, indicating comparable ranking performance, but BiLSTM provides more consistent class-level predictions. These results demonstrate that bidirectional context helps the model better understand subtle sentiment cues, which is particularly useful in longer texts like news articles.

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Mix | 0.67 | 0.40 | 0.50 | 5 |
| Neg | 0.74 | 0.96 | 0.83 | 26 |
| Neu | 1.00 | 0.25 | 0.40 | 4 |
| Pos | 1.00 | 0.50 | 0.67 | 6 |
| Accuracy | - | - | 0.76 | 41 |
| Macro avg | 0.85 | 0.53 | 0.60 | 41 |
| Weighted avg | 0.79 | 0.76 | 0.73 | 41 |

Table 16: Classification report for LSTM.

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Mix | 1.00 | 0.20 | 0.33 | 5 |
| Neg | 0.83 | 0.92 | 0.87 | 26 |
| Neu | 0.75 | 0.75 | 0.75 | 4 |
| Pos | 0.71 | 0.83 | 0.77 | 6 |
| Accuracy | - | - | 0.80 | 41 |
| Macro avg | 0.82 | 0.68 | 0.68 | 41 |
| Weighted avg | 0.82 | 0.80 | 0.78 | 41 |

Table 17: Classification report for BiLSTM.

**BERT**

For this project, a BERT-based model was employed to classify text data into the four sentiment categories: **MIXED, NEGATIVE, NEUTRAL,** and **POSITIVE**. Specifically, the `BertForSequenceClassification` architecture (bert-base-uncased) from the Hugging Face Transformers library was used, which fine-tunes the pre-trained BERT model for downstream classification tasks.

**BERT Architecture Overview:** The model consists of several key components, which work together to encode textual information and produce class predictions:

**Embeddings Layer:**

- **Word Embeddings**: Maps each token in the vocabulary (size 30,522) to a 768-dimensional vector.

- **Position Embeddings**: Adds information about the token's position in the sequence, allowing the model to capture the order of words. This embedding has dimensions of $512 \times 768$.
- **Token Type Embeddings**: Distinguishes between segments in tasks like question answering, with 2 embeddings of size 768.
- **Layer Normalization and Dropout**: Normalizes the embeddings to stabilize training and applies dropout (p = 0.1) for regularization.

**Encoder:** The encoder consists of 12 stacked transformer layers (`BertLayer`), each containing:

- **Self-Attention** (`BertAttention`): Computes contextualized representations for each token using queries, keys, and values, each of size 768. Attention weights are regularized with dropout (p = 0.1).
- **Feedforward Network**: Includes an intermediate dense layer expanding from 768 to 3072 dimensions with a GELU activation, followed by a projection back to 768 dimensions. Each sub-layer is followed by LayerNorm and dropout for stability and regularization. GELU was chosen because it has been shown to be smoother than RELU [12].

**Pooler:** The [CLS] token output is passed through a linear layer followed by a Tanh activation, producing a fixed-size vector summarizing the entire sequence.

**Classification Head:** A dropout layer (p = 0.1) precedes a final linear layer mapping the 768-dimensional pooled output to 4 sentiment classes. During training, cross-entropy loss is used to optimize the classifier.

**BERT Architecture Diagram:**

```
BertForSequenceClassification(
  (bert): BertModel(
    (embeddings): BertEmbeddings(
      (word_embeddings): Embedding(30522, 768, padding_idx=0)
      (position_embeddings): Embedding(512, 768)
      (token_type_embeddings): Embedding(2, 768)
      (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
      (dropout): Dropout(p=0.1, inplace=False)
    )
    (encoder): BertEncoder(
      (layer): ModuleList(
        (0-11): 12 x BertLayer(
          (attention): BertAttention(
            (self): BertSdpaSelfAttention(
              (query): Linear(in_features=768, out_features=768, bias=True)
              (key): Linear(in_features=768, out_features=768, bias=True)
              (value): Linear(in_features=768, out_features=768, bias=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
            (output): BertSelfOutput(
              (dense): Linear(in_features=768, out_features=768, bias=True)
              (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
              (dropout): Dropout(p=0.1, inplace=False)
            )
          )
          (intermediate): BertIntermediate(
            (dense): Linear(in_features=768, out_features=3072, bias=True)
            (intermediate_act_fn): GELUActivation()
          )
          (output): BertOutput(
            (dense): Linear(in_features=3072, out_features=768, bias=True)
            (LayerNorm): LayerNorm((768,), eps=1e-12, elementwise_affine=True)
```

```
            (dropout): Dropout(p=0.1, inplace=False)
          )
        )
      )
    )
    (pooler): BertPooler(
      (dense): Linear(in_features=768, out_features=768, bias=True)
      (activation): Tanh()
    )
  )
  (dropout): Dropout(p=0.1, inplace=False)
  (classifier): Linear(in_features=768, out_features=4, bias=True)
)
```

**Model Fine Tuning Strategies:** In order to enhance the performance of our BERT-based sentiment classification model, we implemented several model tuning strategies. We employed a learning rate warm-up using a linear scheduler to gradually increase the learning rate during the initial training steps, which helped stabilize early training. To address the class imbalance in our dataset, we calculated class weights and incorporated them into the cross-entropy loss function, and additionally performed data augmentation via oversampling of minority classes. We also applied early stopping based on validation loss to prevent overfitting. These strategies collectively contributed to a slight improvement in the model's ability to predict minority classes, as seen in modest gains in precision and recall for categories such as `MIXED` and `NEUTRAL`. However, despite these efforts, the relatively small size of the dataset likely limited the model's capacity to fully learn the patterns associated with the minority classes, resulting in persistent challenges in their accurate prediction.

**Model Performance on the Test Set:** After fine-tuning, the model achieved the following results on the test set of 41 samples:

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| Mix | 1.00 | 0.40 | 0.57 | 5 |
| Neg | 0.71 | 0.85 | 0.77 | 26 |
| Neu | 0.20 | 0.25 | 0.22 | 4 |
| Pos | 1.00 | 0.50 | 0.67 | 6 |
| Accuracy | - | - | 0.68 | 41 |
| Macro avg | 0.73 | 0.50 | 0.56 | 41 |
| Weighted avg | 0.74 | 0.68 | 0.68 | 41 |

Table 18: Classification report for BERT.

**Performance Analysis:**

- The model performs best on **NEGATIVE** sentiment, showing strong recall (0.85), which suggests it is able to correctly identify most negative cases.
- **MIXED** and **POSITIVE** classes have high precision (1.00), meaning when the model predicts these classes, it is usually correct, but recall is lower, indicating some instances are missed.
- The **NEUTRAL** class is challenging, likely due to the small support (4 samples), resulting in very low precision, recall, and F1-score.
- The macro F1-score (0.56) highlights that the model struggles with minority classes, despite overall accuracy being moderate (0.68).

# 5 Conclusion

This project investigated sentiment analysis of youth mental-health–related news articles through three major tasks.

Task A evaluated the performance of multiple large language models (LLMs) in sentiment classification, revealing both consensus and divergence across models. The three LLMs—ChatGPT, Gemini, and Grok—demonstrated strong agreement on the sentiment of the majority of articles, with Gemini achieving the highest overall accuracy and balanced F1-score, Grok excelling in recall, and ChatGPT showing limitations in capturing nuanced sentiment. These results highlight the complementary strengths of LLMs and suggest that combining model predictions could enhance coverage of subtle or underrepresented sentiment classes.

Task B involved a statistical analysis of sentiment class distributions and article characteristics. Negative sentiment predominated across models and human-refined labels, while positive and neutral classes were underrepresented, and mixed sentiment exhibited intermediate frequency and longer article lengths. These findings underscore the importance of addressing class imbalance when training models, as underrepresented sentiments, particularly neutral and positive, may be more challenging to detect accurately. Understanding these distributional issues informed our modeling decisions and helped explain model performance differences.

Task C applied traditional and neural classification models, including logistic regression, LSTM, BiLSTM, and fine-tuned BERT. BiLSTM achieved the highest overall accuracy (80%) and weighted F1-score (0.78), outperforming standard LSTM (accuracy 76%, F1 0.73), fine-tuned BERT (accuracy 68%, F1 0.68), and logistic regression as a classical baseline (accuracy 73%, F1 0.69). Class-level analysis revealed that BiLSTM improved predictions for neutral and positive articles, while BERT achieved high precision for minority classes such as MIXED (1.00) but low recall for NEUTRAL (0.25). Logistic regression provided a reasonable baseline but struggled with underrepresented sentiment classes, highlighting the benefits of sequence modeling and bidirectional context in LSTM-based architectures. Overall, LSTM-based architectures demonstrated robustness in low-resource settings, while transformer models showed potential but required more data for effective fine-tuning.

Across all tasks, key challenges included class imbalance, small dataset size, and the difficulty of capturing nuanced or mixed sentiment in complex news articles. These challenges were addressed through stratified sampling, model comparison across architectures, and qualitative error analysis.

Lessons learned include the importance of aligning model architecture with dataset characteristics, using stratification to ensure fair evaluation, and the complementary strengths of different models for sentiment classification.

Future improvements could involve expanding the dataset, employing advanced class-balancing techniques, leveraging ensemble methods, or experimenting with larger transformer models using parameter-efficient fine-tuning. These steps could enhance performance on minority classes and improve the reliability of sentiment classification in sensitive domains such as youth mental health.

Overall, the combination of semantic analysis, model comparison, and stratified training provides a comprehensive and practical framework for sentiment analysis of social media-related news, offering methodological insights and guidance for future studies.

# 6 Project Management

## 6.1 Project Team

**Joseph Keepers & Kayla DePalma**
*Role: Project Partners*
**Responsibilities and Contributions:**

- **Data Filtering & Annotation:** Both collaborated on cleaning the dataset, removing duplicates and non-English articles. Co-designed prompts for leveraging LLMs to filter and annotate relevant articles, and jointly reviewed and validated the results, resolving any disagreements to finalize the dataset.
- **Data Preparation & Feature Engineering:** Both preprocessed the text data and generated feature representations, including TF-IDF vectors, co-occurrence matrices, and heatmaps.
- **Task A - LLM Comparison:** Both analyzed the performance of LLM models (ChatGPT, Gemini, Grok) against human-refined labels and jointly created tables summarizing results.
- **Task B - Statistical Analysis:** Both performed statistical analysis of sentiment class distributions and article characteristics, producing tables and descriptive summaries.
- **Task C - Model Training and Evaluation:** Both conducted semantic analysis and generated evaluation graphs. Kayla focused on training logistic regression and LSTM models, while Joseph trained and fine-tuned BERT. Both analyzed and interpreted model results collaboratively.
- **Reporting:** Jointly wrote, edited, and compiled the final report.

## References


[1] P. Sood, C. He, D. Gupta, Y. Ning, and P. Wang, "Understanding student sentiment on mental health support in colleges using large language models," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 1865–1872.

[2] A. Samuels and J. Mcgonical, "News sentiment analysis," *ArXiv*, vol. abs/2007.02238, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220364594

[3] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 606–615. [Online]. Available: https://aclanthology.org/D16-1058/

[4] M. Costola, O. Hinz, M. Nofer, and L. Pelizzon, "Machine learning sentiment analysis, covid-19 news and stock market reactions," *Research in International Business and Finance*, vol. 64, p. 101881, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0275531923000077

[5] S. Žitnik, N. Blagus, and M. Bajec, "Target-level sentiment analysis for news articles," *Knowledge-Based Systems*, vol. 249, p. 108939, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095070512200452X

[6] R. Rizal, U. Chandini Pendit, N. Ramli, and S. Annisa, "Sentiment analysis of application x on the impact of social media content on adolescent mental well-being using naive bayes algorithm," *International Journal of Informatics and Computing*, vol. 1, no. 1, pp. 12–18, 2025.

[7] M. E. de Carvalho Souza and L. Weigang, "Grok, gemini, chatgpt and deepseek: Comparison and applications in conversational artificial intelligence," *INTELIGENCIA ARTIFICIAL*, vol. 2, no. 1, 2025.

[8] X. Wu, G. Cai, B. Guo, L. Ma, S. Shao, J. Yu, Y. Zheng, L. Wang, and F. Yang, "A multi-dimensional performance evaluation of large language models in dental implantology: comparison of chatgpt, deepseek, grok, gemini and qwen across diverse clinical scenarios," *BMC Oral Health*, vol. 25, no. 1, p. 1272, 2025.

[9] S. Alaparthi and M. Mishra, "Bert: A sentiment analysis odyssey," *Journal of Marketing Analytics*, vol. 9, no. 2, pp. 118–126, 2021.

[10] A. Bello, S.-C. Ng, and M.-F. Leung, "A bert framework to sentiment analysis of tweets," *Sensors*, vol. 23, no. 1, p. 506, 2023.

[11] E. Biswas, M. E. Karabulut, L. Pollock, and K. Vijay-Shanker, "Achieving reliable sentiment analysis in the software engineering domain using bert," in *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 2020, pp. 162–173.

[12] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2016, arXiv preprint, version 5, updated 2023. [Online]. Available: https://arxiv.org/abs/1606.08415

# A Appendix: Additional Figures

| yt or sc | title | body_text | provider | publication_date | sentiment_ | sentiment_c | sentiment_gemini | kayla_rating | joe_rating | validated_rating |
|---|---|---|---|---|---|---|---|---|---|---|
| yt | YouTube urged to promote 'high-quality' children's TV | LONDON – The government has urged video platforms like Y | {'name': 'nation.com.pk', 'domain': 'nation.com.pk'} | 2024-12-31T00:00:00+00:0 | MIXED | MIXED | MIXED | MIXED | MIXED | MIXED |
| yt | Steven He: How 'Emotional damage' guy hit 2 billion YouTube views | Chinese-Irish YouTuber and media personality Steven He, com | {'name': 'jang.com.pk', 'domain': 'jang.com.pk'} | 2024-12-31T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | YouTube must do more to educate children, says culture secretary | YouTube must do more to promote children's education and c | {'name': 'The Times', 'domain': 'thetimes.co.uk'} | 2024-12-30T00:00:00+00:0 | MIXED | MIXED | MIXED | MIXED | MIXED | MIXED |
| yt | Laiszve Dambuza admits himself into a mental health wellness centre | Renowned YouTube creator and reality television personality L | {'name': 'iharare.com', 'domain': 'iharare.com'} | 2024-12-18T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | YouTube to turn eligible doctors' channels into official health info sources | YouTube previously worked with accredited health organisatio | {'name': 'The Hindu', 'domain': 'thehindu.com'} | 2024-12-17T00:30:32+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Nearly half of US teens are online 'constantly,' Pew report finds - TheGrio | Nearly half of American teenagers say they are online "consta | {'name': 'thegrio.com', 'domain': 'thegrio.com'} | 2024-12-15T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | Social Media's Grip on Teen Lives: A Deep Dive \| Headlines | A recent report by the Pew Research Centre highlights the pe | {'name': 'devdiscourse.com', 'domain': 'devdiscourse.com'} | 2024-12-12T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | YouTube launches new feature for registered health professionals to reach people | YouTube launches new feature for registered health profession | {'name': 'Telangana Today', 'domain': 'telanganatoday.com'} | 2024-12-10T10:05:15+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | YouTube New Feature Launched To Help Indian Viewers Find High-Quality Health Information Fa | New Delhi, December 10: Google-owned YouTube on Tuesday | {'name': 'latestly.com', 'domain': 'latestly.com'} | 2024-12-10T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | YouTube Announces New Feature For Registered Health Professionals In India -All You Want To | New Delhi: Google-owned YouTube on Tuesday announced n | {'name': 'zeenews.india.com', 'domain': 'zeenews.india.com'} | 2024-12-10T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Ex-Love Island star reveals changes he didn't expect two years after quitting alcohol | In a nation where a pint at the pub is part of our cultural identi | {'name': 'Daily Mail Online', 'domain': 'dailymail.co.uk'} | 2024-12-08T18:13:43+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Rise of child influencers in Chandigarh: Fun hobby or cause for concern? - Hindustan Times | The trend of young children becoming social media influencer | {'name': 'hindustantimes.com', 'domain': 'hindustantimes.com'} | 2024-11-23T00:00:00+00:0 | MIXED | MIXED | MIXED | MIXED | MIXED | MIXED |
| yt | National PTA Awards Over $1M to PTAs to Support Digital Safety, Healthy Lifestyles, Arts and S | ALEXANDRIA, Va., Nov. 20, 2024 /PRNewswire/ -- National PT | {'name': 'Cision PR Newswire', 'domain': 'prnewswire.com'} | 2024-11-20T18:30:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Netizens Ask Samay Raina To 'Start OnlyFans' After His Ad Revenue Comment On Deepika Pad | Netizens Ask Samay Raina To 'Start OnlyFans' After His Ad Re | {'name': 'timesnownews.com', 'domain': 'timesnownews.com'} | 2024-11-19T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Why Joking About Deepika Padukone's Depression on India's Got Latent Wasn't Just 'Dark Hur | In the latest episode of India's Got Latent, Samay Raina's pop | {'name': 'scoopwhoop.com', 'domain': 'scoopwhoop.com'} | 2024-11-18T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Mind Reading: Michael D. Ratner's Lens On Mental Health Storytelling In Pop Culture | Demi Lovato presented Michael D. Ratner with the Project He | {'name': 'Forbes', 'domain': 'forbes.com'} | 2024-11-14T22:20:21+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Analysis: Should India Ban Social Media For Under-16s, Like Australia? | Studies have found that unlimited screen time can disrupt slee | {'name': 'ndtv.com', 'domain': 'ndtv.com'} | 2024-11-14T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | The Social Conundrum: Navigating the Social Media Impact on Youth \| Headlines | The intensifying debate over social media's impact on youth h | {'name': 'devdiscourse.com', 'domain': 'devdiscourse.com'} | 2024-11-09T00:00:00+00:0 | MIXED | MIXED | MIXED | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Australia wants a total social media ban for teens under 16, targeting even YouTube - PhoneAre | Australia is moving forward with a hard-line policy to ban soci | {'name': 'phonearena.com', 'domain': 'phonearena.com'} | 2024-11-07T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Australia proposes social media ban for everyone under the age of 16, citing concerns about yo | Social media hasn't exactly been a net positive for kids and te | {'name': 'pcgamer.com', 'domain': 'pcgamer.com'} | 2024-11-07T00:00:00+00:0 | MIXED | MIXED | MIXED | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Mental Health Student Athletes Justaroarnaway Com - Ponasa | helping student athletes put their mental health first youtube S | {'name': 'ponasa.condesan-ecoandes.org', 'domain': 'ponasa.cc | 2024-11-01T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | European Commission requests for information about recommender systems under the Digital S | The European Commission has sent a request for information | {'name': 'scl.org', 'domain': 'scl.org'} | 2024-10-31T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | 'Being true to yourself is key': Prajakta Koli discusses the importance of mental health with Ansr | Ananya Panday recently ventured into the world of podcasting | {'name': 'Hindustan Times', 'domain': 'hindustantimes.com'} | 2024-10-21T10:02:49+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | 6 tribes have filed lawsuits against social media companies for teen mental health concerns | With Standing Rock's recent filing, there are now six Indigeno | {'name': 'The Bismarck Tribune', 'domain': 'bismarcktribune.com' | 2024-10-21T11:15:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Woman discovers 'secret' note from her late father is his final Dad Joke - Upworthy | A recent Reddit thread (rarr!) that posed the theory: "Gen Z c | {'name': 'upworthy.com', 'domain': 'upworthy.com'} | 2024-10-19T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Lawsuit alleges social media giants exacerbating Native American youth mental health crisis | Social media giants sued over worsening Native American you | {'name': 'CBS News', 'domain': 'cbsnews.com'} | 2024-10-14T14:05:15+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Social media's role in mental health and digital wellbeing \| The Daily Star | As digital platforms become more integral to daily life, concer | {'name': 'thedailystar.net', 'domain': 'thedailystar.net'} | 2024-10-13T00:00:00+00:0 | MIXED | MIXED | MIXED | POSITIVE | POSITIVE | POSITIVE |
| yt | I'm a psychiatrist and here are 5 signs you might have undiagnosed PTSD \| Daily Mail Online | A psychiatrist has revealed five subtle signs that suggest you | {'name': 'dailymail.co.uk', 'domain': 'dailymail.co.uk'} | 2024-10-12T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | Trae Young And Iman Shumpert Talk Mental Health In Invisible Game A New Unscripted Series F | It's that time of the year again when we elucidate mental healt | {'name': 'essence.com', 'domain': 'essence.com'} | 2024-10-10T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | In Honor of World Mental Health Day, Ellie Thumann Shares a Grounded Day in Her Life | Ellie Thumann is best known for her lifestyle, beauty, and fash | {'name': 'swimsuit.si.com', 'domain': 'swimsuit.si.com'} | 2024-10-10T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | Trae Young, Percy 'Master P' Miller, Iman Shumpert Talk Mental Health in INVISIBLE GAME, a N | UNINTERRUPTED, the award-winning athlete storytelling bran | {'name': 'businesswire.com', 'domain': 'businesswire.com'} | 2024-10-08T00:00:00+00:0 | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE | POSITIVE |
| yt | TikTok is designed to be addictive to kids and causes them harm, US states' lawsuits say | "They've chosen profit over the health and safety, well-being o | {'name': 'qz.com', 'domain': 'qz.com'} | 2024-10-08T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Ananya Panday opens up on finding it hard to express herself while feeling low; 'I start crying im | "I find it tough to express myself," Ananya Panday said during | {'name': 'pinkvilla.com', 'domain': 'pinkvilla.com'} | 2024-10-07T00:00:00+00:0 | MIXED | MIXED | MIXED | MIXED | MIXED | MIXED |
| yt | The EU wants to know more about the content 'rabbit holes' and the way TikTok, YouTube, and | The EU Commission asked YouTube, Snapchat, and TikTok to | {'name': 'phonearena.com', 'domain': 'phonearena.com'} | 2024-10-03T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | EU requests info from YouTube, Snapchat, TikTok on content algorithms | YouTube and TikTok apps are seen on a smartphone in this illu | {'name': 'The Indian Express', 'domain': 'indianexpress.com'} | 2024-10-03T03:06:22+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | YouTube, TikTok And Snapchat Face EU Scrutiny Over Algorithmic Role In Spreading Harmful C | The EU Commission made a request for data from YouTube, S | {'name': 'Jagran English', 'domain': 'english.jagran.com'} | 2024-10-03T06:14:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | Arkansas Lawsuit: Google's YouTube Fuels Youth Mental Health Crisis | Arkansas has filed a lawsuit against YouTube and its parent co | {'name': 'breitbart.com', 'domain': 'breitbart.com'} | 2024-10-02T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | EU demands information from YouTube, Snapchat, TikTok on content algorithms \| Tech News - E | EU demands information from YouTube, Snapchat, TikTok on | {'name': 'business-standard.com', 'domain': 'business-standard | 2024-10-02T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | Lawmakers Demand Mental Health Warning Labels on Social Media | After a slew of claims that social media is destroying mental h | {'name': 'vice.com', 'domain': 'vice.com'} | 2024-10-02T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | EU requests information from YouTube, Snapchat, TikTok about recommendations - SWI swissi | This content was published on October 2, 2024 - 12:21 PARIS | {'name': 'swissinfo.ch', 'domain': 'swissinfo.ch'} | 2024-10-02T00:00:00+00:0 | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL | NEUTRAL |
| yt | Children would lie about their age to evade YouTube ban | Preventing Australian teenagers from accessing YouTube woul | {'name': 'thewest.com.au', 'domain': 'thewest.com.au'} | 2024-10-02T00:00:00+00:0 | MIXED | MIXED | MIXED | MIXED | MIXED | MIXED |
| yt | Arkansas Sues YouTube, Alphabet Alleging Harms To Kids' Mental Health - KFF Health News | Arkansas Sues YouTube, Alphabet Alleging Harms To Kids' M | {'name': 'kffhealthnews.org', 'domain': 'kffhealthnews.org'} | 2024-10-01T00:00:00+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |
| yt | Lawsuit accuses YouTube of harming youth mental health | Arkansas sued YouTube and parent company Alphabet on Mo | {'name': 'Fortune', 'domain': 'fortune.com'} | 2024-10-01T00:46:41+00:0 | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE | NEGATIVE |

Figure 1: Finalized Dataset: 401 articles

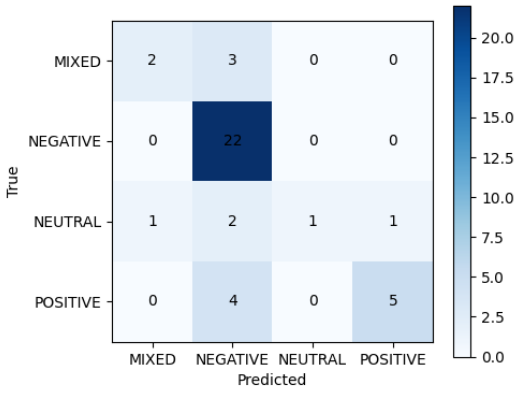


Figure 2: Confusion Matrix: Logistic Regression without Stratified Sampling

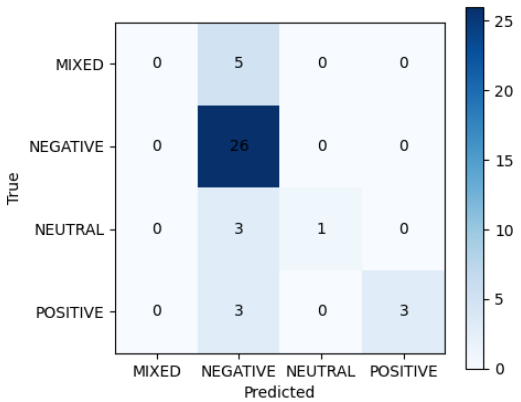


Figure 3: Confusion Matrix: Logistic Regression with Stratified Sampling

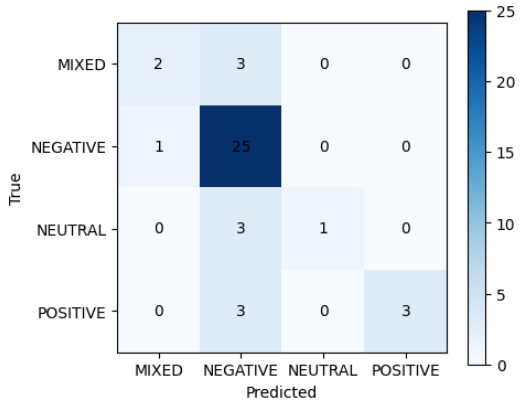


Figure 4: Confusion Matrix: LSTM Standard

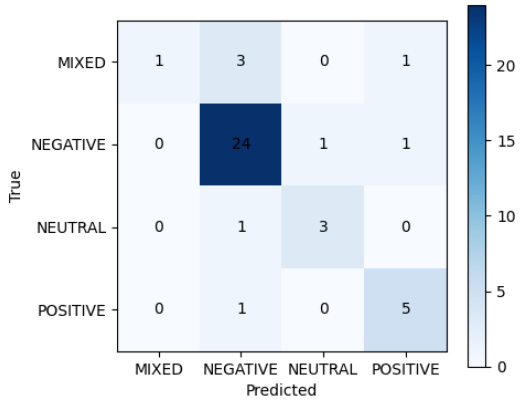


Figure 5: Confusion Matrix: LSTM Bi-Directional

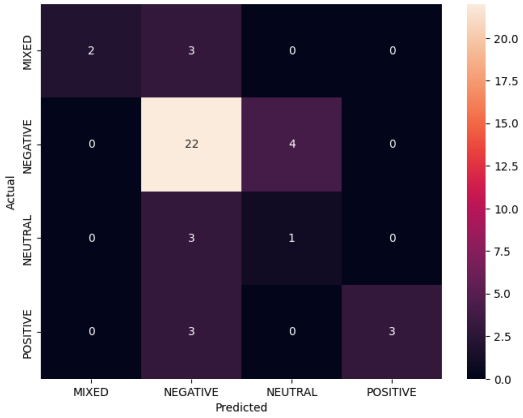


Figure 6: Confusion Matrix: pre-trained BERT

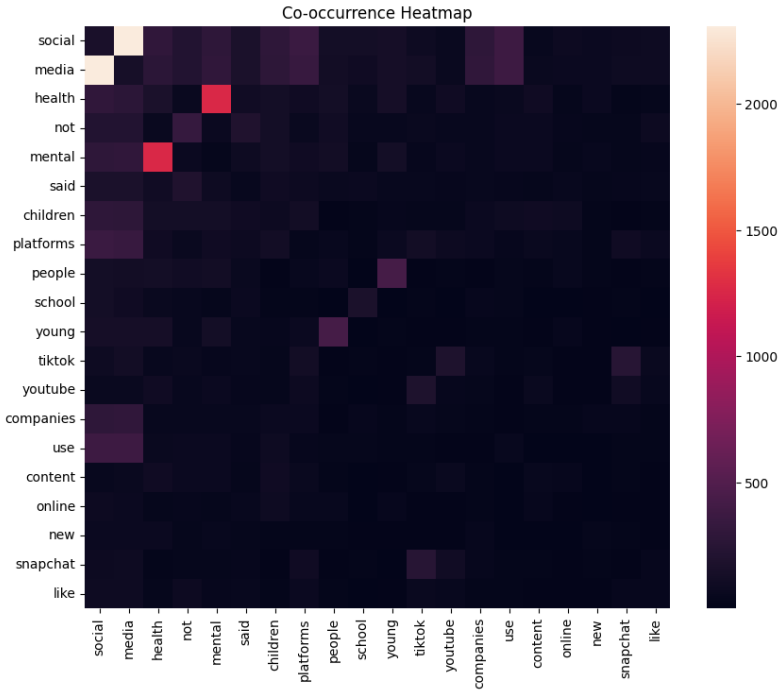


Figure 7: Heat Map: Top 20 Related Words

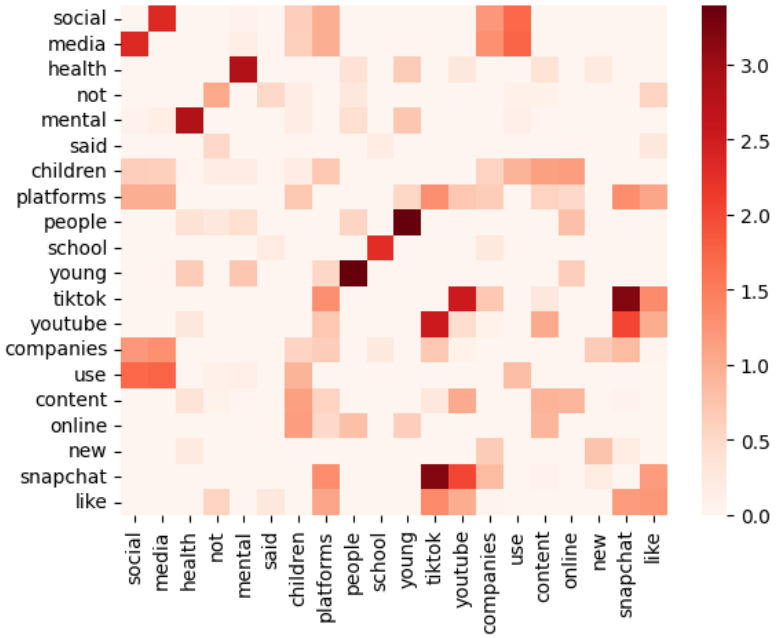


Figure 8: Heat Map: Top 20 Related Words with PPMI

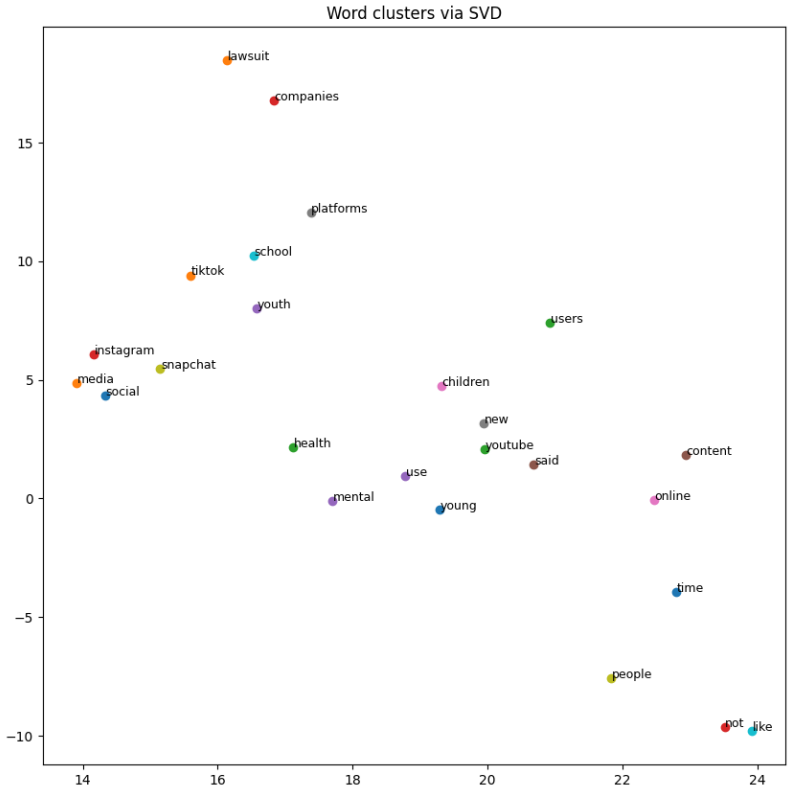


Figure 9: Word Clusters: SVD

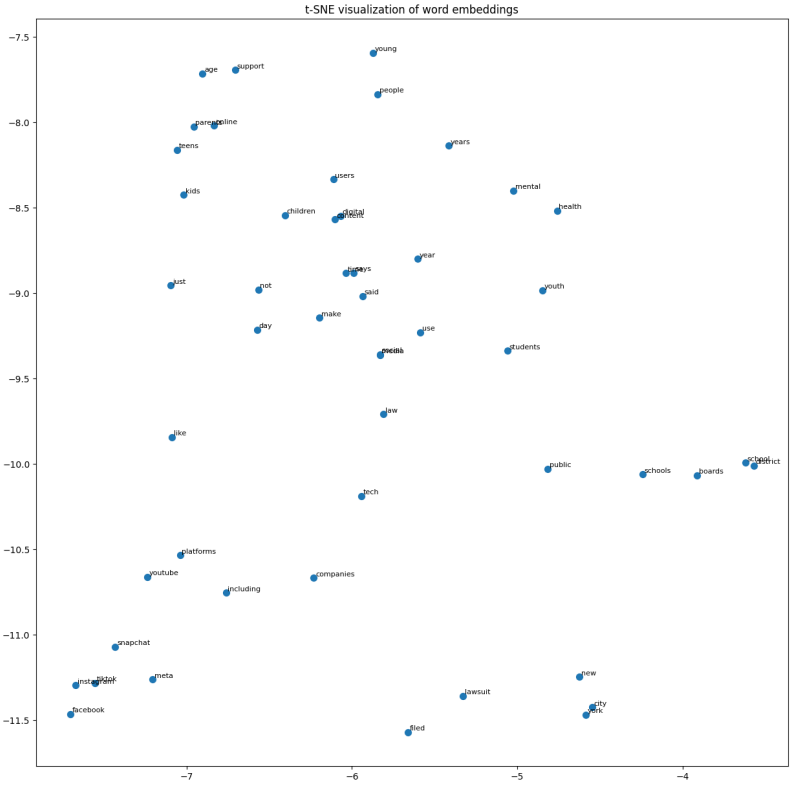


Figure 10: Word Clusters: t-SNE word-embeddings