Reddit Thread Summarizer

Inside Threads: Where Data Meets Dialogue

Where data meets dialogue, we summarize what matters!

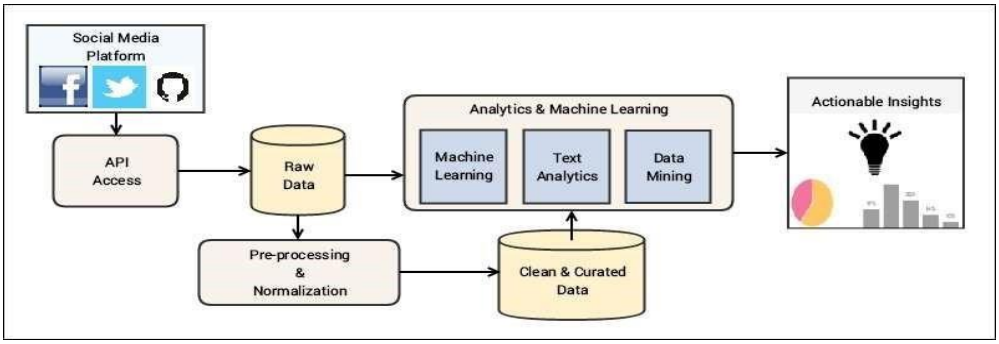
Report: Reddit Thread Summarizer

Overview

Social media analytics (SMA) involves collecting and analyzing data from social media platforms to understand user behavior, engagement, and trends. By utilizing various tools, organizations can track metrics like engagement, reach, sentiment, and demographics, which helps optimize their social media strategies. This process includes monitoring conversations and trends to evaluate campaign effectiveness and identify opportunities and challenges.

Key objectives of social media analytics include:

- Understanding Audience Behavior: Sentiment analysis (e.g., VADER, Text Blob) classifies opinions about brands and topics as positive, negative, or neutral by analyzing large text data volumes.
- Topic Detection: Techniques like Latent Dirichlet Allocation reveal trending themes in social conversations, indicating what content resonates with audiences.
- Influencer Detection: Network analysis identifies influential users and brand advocates based on engagement metrics, aiding targeted marketing efforts.
- Trend Prediction: Time Series analysis of historical data forecasts future trends and viral behaviors, allowing brands to engage with emerging topics promptly.
- Customer Segmentation: Clustering algorithms utilize social media conversations to accurately segment users into personas based on preferences.
- Post Engagement Prediction: Predictive modeling benchmarks the potential success of new content based on historical performance metrics like hashtags and timing.
- Privacy Risk Identification: Pattern recognition and keyword filters scan user data to detect potential privacy violations or non-compliance with regulations like GDPR.



Problem Statement

The objective of this project on Reddit Post Summarization centers on the challenge of efficiently summarizing information from an extensive volume of user-generated content on the platform. With approximately 469 million posts published in 2023 and an average of 7.5 million comments generated daily, distilling this extensive array of information into concise summaries presents significant challenges. Each post may encompass diverse opinions, nuanced arguments and varying levels of content, making it difficult to extract key themes and sentiments without losing critical information.

Here are the primary problems this project aims to tackle:

- **Long-Form Content Navigation:** Reddit threads often contain thousands of comments, making it

challenging for users to find valuable information. A summarizer can highlight critical comments, extract main themes, and streamline user navigation through lengthy discussions.

- **Repetitive Information:** Users frequently repeat similar points in threads. Filtering out these repetitive comments can provide unique perspectives and clarify information. The summarizer can quantify the frequency of responses, helping users grasp collective opinions while highlighting nuanced contributions.
- **Context Loss:** In lengthy discussions, context may be lost as comments diverge into tangents. A summarizer can address this by providing necessary background information, ensuring coherence and helping users follow main points without losing the overall narrative. It will preserve the integrity of complex discussions while making them easier to understand.
- **Trending Topics Identification:** Users want to stay updated on hot topics but often struggle to sift through vast content. The summarizer can extract the most discussed and upvoted themes, offering a curated snapshot of trending topics in real time.
- **Sentiment Analysis:** Incorporating sentiment analysis would give users insight into the thread's tone (positive, negative, or neutral). By identifying emotional highs and lows, the summarizer can help users navigate volatile discussions and make informed engagement decisions.
- **Integration with Other Tools:** The summarizer could integrate with social media platforms, notetaking apps, or project management software, allowing users to save and organize insights efficiently. Integration with browser extensions would enable users to access summarized content seamlessly while browsing, enhancing the Reddit experience.

This project addresses these challenges by developing a summarization model tailored to Reddit's unique structure, enhancing user engagement and providing actionable insights.

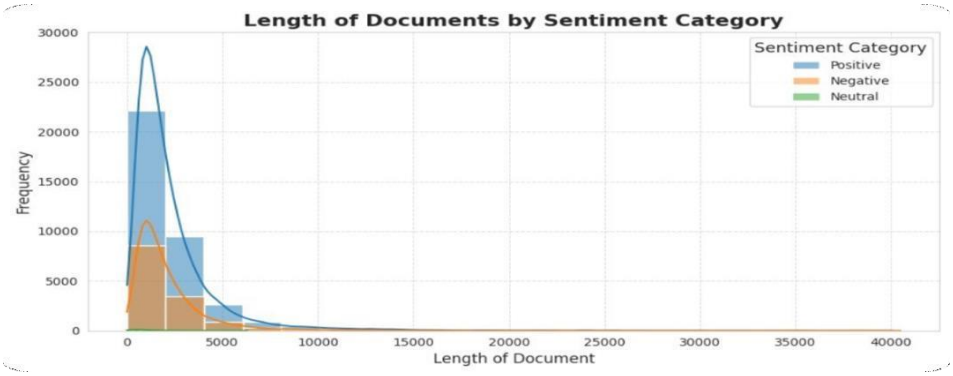
Data Collection and Understanding

The project utilized the TLDRHQ dataset, a comprehensive resource designed for summarization tasks. The dataset consists of the following key components:

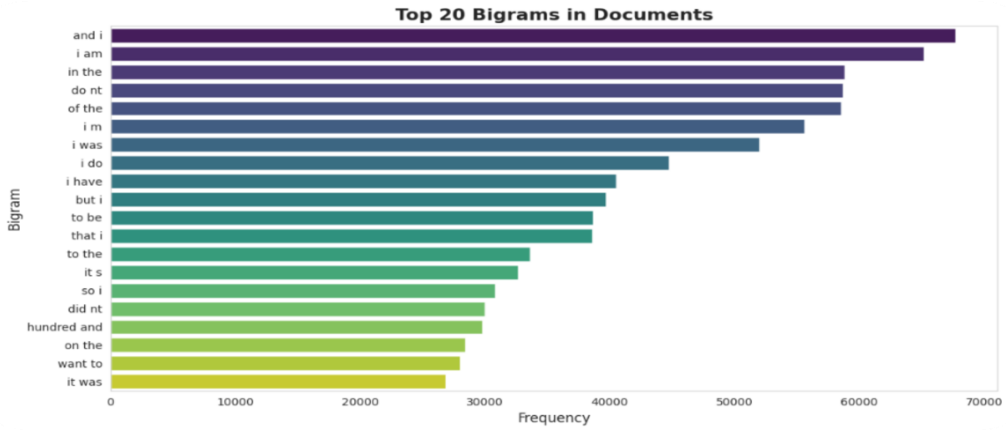
- **Id**: Unique identifiers for posts, generated to differentiate submissions (RS) from comments (RC).
- **Document**: User posts, segmented by sentences, with special tokens marking boundaries.
- **Summary**: Concise, user-generated descriptions (TL; DR) summarizing the content of posts.
- **Ext_labels**: Extractive labels, highlighting significant sentences within the document.
- **Rg_labels**: Rouge scores, quantifying sentence alignment with the summary based on Rouge-2 and Rouge-L metrics.

Insights Derived from the Dataset

- Positive sentiment documents tend to be more detailed and longer, reflecting elaborate responses.



- Linguistic patterns emphasize personal expressions, such as frequent use of first-person narratives and conversational tones.



Data Preprocessing

The analysis was conducted using Databricks in a distributed cluster environment to ensure efficient handling of large datasets. To optimize the dataset for summarization, the following steps were implemented:

1. **Data Cleaning**
   - Removed unnecessary elements such as URLs, HTML tags, email addresses, and usernames (e.g., "@username" and "u/username") using regular expressions.
   - Stripped out excessive spaces and normalized text formats.
2. **Text Normalization**
   - Converted all text to lowercase for uniformity.
   - Preserved punctuation and stop words to maintain the natural flow and context of Reddit's informal language.
3. **Handling Slangs**
   - Replaced slang terms and abbreviations with their standard equivalents using a comprehensive dictionary. For instance, "u" was replaced with "you," and "lol" with "laughing out loud."
4. **Feature Extraction**
   - Transformed array-based features (e.g., ext_labels and rg_labels) into string format for compatibility with downstream tasks.
5. **Sentiment Analysis**
   - Applied VADER (Valence Aware Dictionary and sentiment Reasoner) for sentiment scoring, categorizing content as Positive, Negative, or Neutral based on compound scores.
6. **Tokenization and Topic Modeling**
   - Tokenized text into individual words or phrases.
   - Leveraged Latent Dirichlet Allocation (LDA) for topic modeling, identifying predominant themes within posts and summaries.

These preprocessing steps ensured the data was clean, structured, and ready for training an abstractive summarization model tailored to Reddit's unique content dynamics.

Model Selection and Training

Why BART for Text Summarization:

- **Contextual Understanding:** BART captures both left and right context in a sentence, which helps in better understanding of the text.
- **Fine-tuning Capability:** BART can be fine-tuned for specific tasks like summarization, making it adaptable and effective for generating coherent, human-like summaries.
- BART was chosen for its ability to combine a bidirectional encoder (understanding context from both directions) with an autoregressive decoder (generating fluent summaries). Its pretraining as a denoising autoencoder makes it robust for handling noisy Reddit threads. BART excels in handling long texts, maintaining context, and generating coherent, context-aware summaries. Its fine-tuning flexibility, ability to freeze layers for efficiency, and strong performance on benchmarks like CNN/Daily Mail make it ideal for this project. Additionally, it effectively processes informal and unstructured data, ensuring high-quality outputs for Reddit-specific summarization tasks.

Initial Implementation of BART model & Evaluation

What We Did:

Trained BART without freezing layers to allow full model learning and monitored training and validation losses to detect overfitting/underfitting.

Why We Did It:

Baseline Evaluation: Established baseline performance for comparison with future fine-tuning.

- Training Loss decreased, indicating improvement.
- Validation Loss fluctuated, suggesting potential overfitting.

Epoch	Training Loss	Validation Loss
1	1.078400	0.989186
2	0.761700	0.989391
3	0.536200	1.063200

Fine-Tuning BART model using Transfer Learning & Model Evaluation

Freezing Layers:

Strategy: Freezing certain layers (e.g., encoders and decoders) helps retain pre-trained knowledge while adapting the model to new tasks.

Impact: Freezing allows faster convergence and reduces the risk of overfitting, focusing learning on specific layers.

Learning Rate Adjustment:

Learning Rates Used: Tested 5e-5 and 1e-4 learning rates to optimize model performance.

5e-5: Slower learning, more gradual adjustments, reduces the risk of overshooting.

1e-4: Faster learning, potentially better for capturing larger patterns but may lead to overfitting if too high.

To enhance the summarization process, we fine-tuned the BART model using transfer learning, experimenting with freezing encoder and decoder layers. Key configurations and their observations included:

5 encoder and 5 decoder layers

Epoch	Training Loss	Validation Loss
1	0.358100	0.360509
2	0.292200	0.345791
3	0.269500	0.350126

6 encoder and 6 decoder layers

Epoch	Training Loss	Validation Loss
1	0.455100	0.378671
2	0.259200	0.361149
3	0.322400	0.361094

6 encoders and 3 decoders

Epoch	Training Loss	Validation Loss
1	1.068900	0.986812
2	0.763000	0.983820
3	0.554600	1.059018

7 encoders and 7 decoders

Epoch	Training Loss	Validation Loss
1	0.304800	0.352052
2	0.311500	0.340560
3	0.279400	0.344072

- **5 Encoders, 5 Decoders (5E-5D):**
  - Training Loss decreased, indicating learning progress.
  - Validation Loss fluctuated, revealing challenges in generalization.
  - Achieved the best ROUGE-L (fluency), but with lower BLEU (precision).
- **6 Encoders, 6 Decoders (6E-6D):**
  - Training Loss showed consistent improvement.
  - Validation Loss remained stable, suggesting better generalization.
  - ROUGE scores were slightly lower than 5E-5D.
- **6 Encoders, 3 Decoders (6E-3D):**
  - Balanced performance with high ROUGE-1, ROUGE-2, and decent BLEU scores.
  - Validation Loss fluctuated, indicating room for improvement.
- **7 Encoders, 7 Decoders (7E-7D):**
  - Training Loss steadily decreased.
  - Validation Loss was stable but resulted in the lowest BLEU and weaker ROUGE scores.

Challenges and Observations:

While these configurations provided valuable insights, issues such as **prolonged training times and overly neutral summaries prompted further refinements.** Evaluation metrics (ROUGE, BLEU) and validation loss trends indicated potential overfitting or suboptimal generalization.

Advanced Fine-Tuning

To address these challenges, additional configurations were explored:

Optimization Techniques

1. Introduced **mixed precision training** to handle high training times efficiently.
2. Adjusted summary length to reduce neutral tones and improve contextual relevance.

- **9 Encoders, 9 Decoders (9E-9D):**
  - Training Loss steadily decreased.
  - Validation Loss showed slight increases after Epoch 3, indicating potential overfitting.

Epoch	Training Loss	Validation Loss
1	1.049700	0.972282
2	0.836000	0.966887
3	0.700600	1.005086

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
Average ROUGE-1 Score (shorter): 0.2626
Average ROUGE-2 Score (shorter): 0.0728
Average ROUGE-L Score (shorter): 0.1738
Average BLEU Score (shorter): 0.0215
Average BERT F1 Score (shorter): 0.2228

Start coding or generate with AI.
```

- **10 Encoders, 10 Decoders (10E-10D):**
  - Training Loss improved consistently.
  - Validation Loss stabilized but remained higher, suggesting suboptimal training.

Epoch	Training Loss	Validation Loss
1	1.449300	1.377063
2	1.288100	1.351416
3	1.182800	1.359201

- **11 Encoders, 11 Decoders (11E-11D):**
  - Training Loss decreased steadily.
  - Validation Loss showed consistent decreases across epochs with no increases.
  - Demonstrated superior performance, balancing training and validation metrics effectively.

Epoch	Training Loss	Validation Loss
1	0.983200	0.965822
2	0.901700	0.957154
3	0.837800	0.958458

config.json: 100%

```
Average ROUGE-1 Score: 0.2623
Average ROUGE-2 Score: 0.0755
Average ROUGE-L Score: 0.1742
Average BLEU Score: 0.0243
Average BERT F1 Score: 0.2201
```

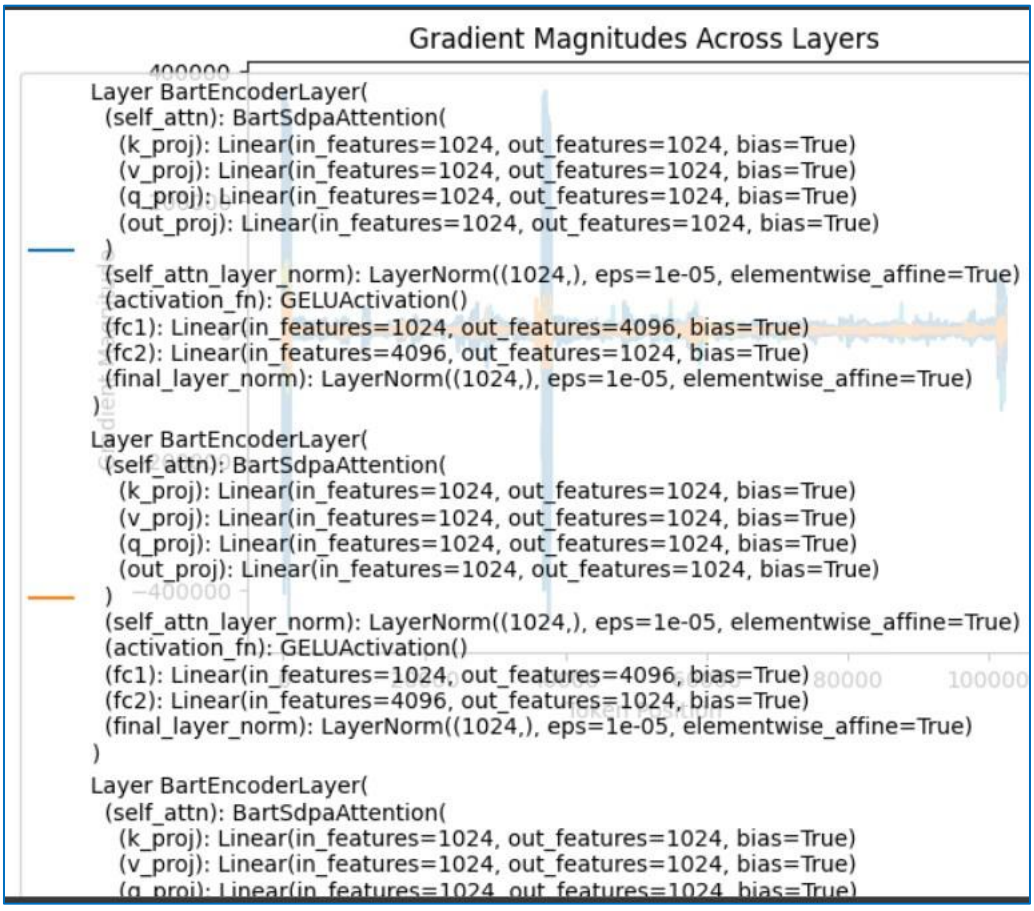
Final Recommendation The 11E-11D configuration emerged as the most promising, this setup balanced fluency, precision, and computational efficiency, providing a robust solution for Reddit thread summarization.

- Consistently decreasing validation loss.
- Improved evaluation metrics, including BERT F1 scores.

Model Interpretability

Low Gradient Layers: Early encoder layers have low gradients, suggesting they capture general pre-trained knowledge and can be frozen to save computation.

High Gradient Layers: Later encoder layers and initial decoder layers show higher gradients, indicating they adapt more to the task and should remain trainable.



Incorporating RAKE with BART for Summarization

To enhance the summarization process, **Rapid Automatic Keyword Extraction (RAKE)** was used to identify critical keywords from Reddit threads before feeding the data into the BART model. Here's how RAKE contributed:

1. **Keyword Extraction**:
    a. RAKE efficiently extracted important keywords from the text by analyzing word co-occurrences and their positions within the document.
    b. These keywords represented the most critical aspects of the document, helping to focus on the core themes.
2. **Integration with BART**:
    a. The extracted keywords were combined with the document text, providing enriched input for the BART model.
    b. This helped the model focus on the most relevant information during the summarization process.
3. **Benefits**:
    a. **Improved Relevance**: By highlighting key information, RAKE helped BART generate summaries that were more aligned with the document's main points.
    b. **Noise Reduction**: Keywords acted as a filter to reduce the influence of irrelevant or redundant information.
    c. **Efficiency**: The use of RAKE ensured quick and effective preprocessing, complementing BART's summarization capabilities.

This integration allowed the BART model to produce concise and contextually rich summaries by leveraging the key insights extracted by RAKE.

Incorporating Topics Using LDA with BART

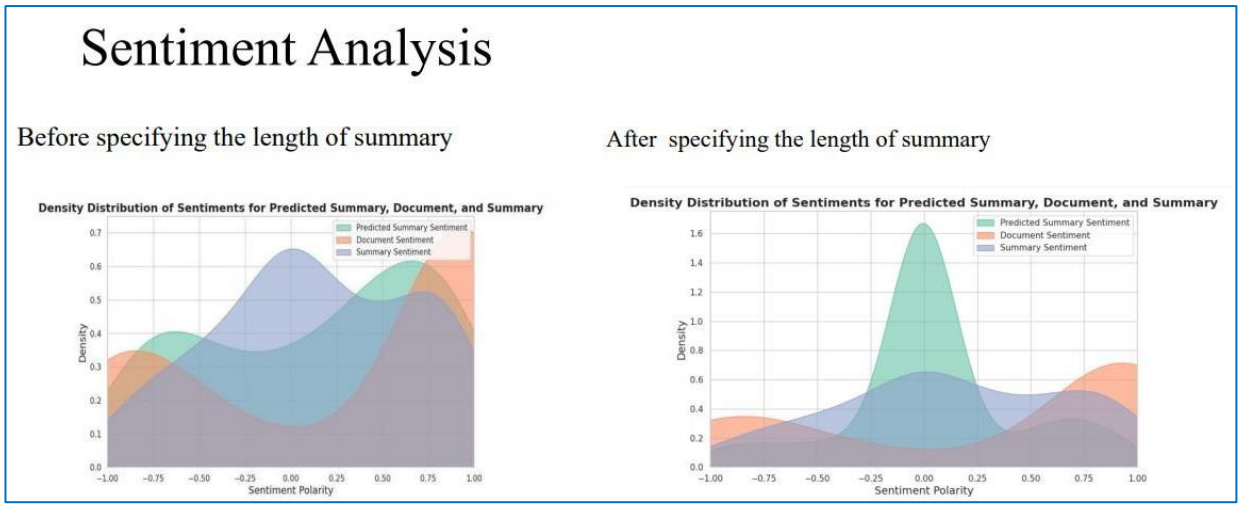
To enhance the summarization process, Latent Dirichlet Allocation (LDA) was used to extract topics from the Reddit threads. This step adds contextual information to the input data, which helps the BART model generate summaries that are more aligned with the underlying themes. Here's how this integration works:

1. **Topic Extraction**:
    a. LDA identifies dominant topics in the text by analyzing word co-occurrence patterns.
    b. Each document is assigned a topic distribution, providing a structured representation of its key themes.
2. **Combining Topics with Documents**:
    a. The extracted topic information is concatenated with the original document text. This enriched input provides additional context to the BART model, enabling it to generate more relevant and context-aware summaries.
3. **Benefits**:
    a. **Improved Context Understanding**: By including thematic information, BART captures deeper semantic relationships, producing summaries that are more coherent and representative of the document's core ideas.
    b. **Reduced Noise**: LDA helps filter out less relevant details, focusing the summarization process on the most critical aspects.

c. **Enhanced Personalization**: Topics provide a high-level summary of the content, aiding in creating summaries tailored to specific user needs or interests.

This combination of LDA and BART leverages the strengths of both topic modeling and deep learning to deliver high-quality, context-aware summarization.

Data Visualizations and Analysis



The visualizations for sentiment analysis compare the sentiment scores of documents before and after specifying the length of the summary. This comparison provides insight into how summary constraints influence the overall sentiment derived from the text.

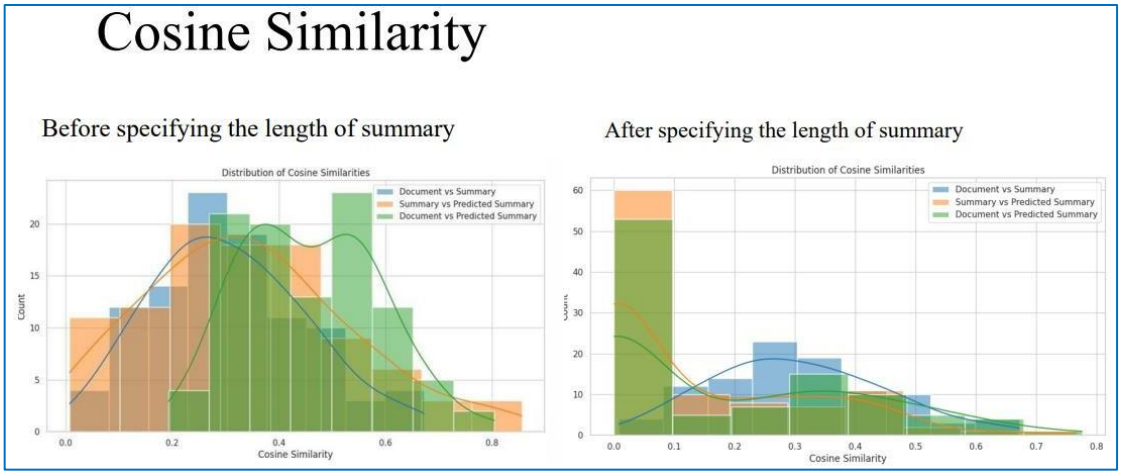
Before Specifying Summary Length:

- The sentiment analysis visual depicts a wide variance in sentiment scores, indicating raw text sentiment without length normalization.
- Some documents show extreme sentiment values (highly positive or negative), reflecting their unfiltered, original emotional tone.

After Specifying Summary Length:

- The sentiment scores appear more centralized, with fewer extremes. This suggests that constraining the summary length leads to a more neutralized or uniform sentiment distribution.
- The neutralization likely arises because shorter summaries may omit emotionally charged language, focusing on key content instead.

Key Observation: The reduction in sentiment extremes after specifying the summary length indicates that sentiment is sensitive to text compression and abstraction. This finding could be important for applications requiring balanced emotional representation.



Cosine similarity visualizations illustrate how the similarity between documents changes before and after specifying the length of their summaries. These insights are critical for understanding text relevance and semantic alignment.

Before Specifying Summary Length:

- High variability in cosine similarity scores suggests that longer texts contain more unique terms or divergent information.
- The distribution highlights instances of both high and low semantic overlap across document pairs.

After Specifying Summary Length:

- The cosine similarity scores converge to a narrower range, reflecting improved consistency in semantic content when summaries are constrained in length.
- Shorter summaries likely emphasize core concepts, reducing noise and enhancing comparability.

Key Observation: The shift toward more consistent cosine similarity values underscores the role of summary length in standardizing semantic representation. This finding is particularly valuable for tasks like clustering, where uniform document similarity improves grouping accuracy.

Visual Implications

- For Sentiment Analysis: Adjusting summary length can strategically neutralize or amplify sentiment, depending on application needs (e.g., customer feedback analysis, automated content creation).
- For Cosine Similarity: Standardizing summary lengths improves the reliability of similarity measures, critical for machine learning models relying on text clustering or classification.

Enhancing Text Summarization Using Topic Modeling and BART

This code combines Latent Dirichlet Allocation (LDA) for topic modeling and a BART (Bidirectional and Auto-Regressive Transformer) model for text summarization. The workflow involves preprocessing, extracting topics, appending topic-specific keywords to documents, and training a BART model on the augmented dataset.

Topic Augmentation with LDA

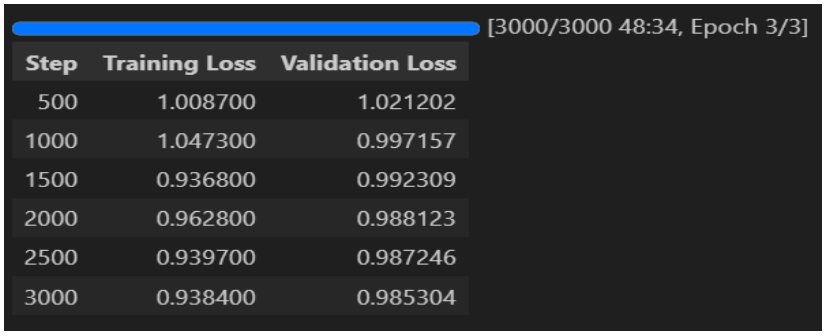
- Extracts key topics from documents using Latent Dirichlet Allocation (LDA).
- Appends top topic keywords to each document, enriching them with critical contextual information.
- Ensures the summarization model captures topic-specific nuances, improving relevance and detail in generated summaries.

Fine-Tuning BART for Summarization

- Leverages BART, a pre-trained transformer model, for generating high-quality summaries.
- Fine-tuned on documents enhanced with topic keywords, enabling the model to focus on important information.
- Freezes selective layers of the model, preserving general language understanding while adapting to the specific dataset.

Benefits of the Approach

- Improved Contextual Understanding: Enriching documents with topic-specific keywords enhances the model's ability to generate summaries that align with the core themes.
- Efficient Processing: Selective fine-tuning and advanced training techniques optimize performance without sacrificing quality.
- Scalability: The combined approach of LDA and BART scales effectively across large datasets, enabling robust summarization for varied topics.



Step	Training Loss	Validation Loss
500	1.008700	1.021202
1000	1.047300	0.997157
1500	0.936800	0.992309
2000	0.962800	0.988123
2500	0.939700	0.987246
3000	0.938400	0.985304

Key Observations:

- Both losses decrease gradually, indicating effective learning.
- By the final step (3000), training loss is 0.9384, and validation loss is 0.9853. The small gap suggests the model balances training accuracy with generalization.

Evaluation

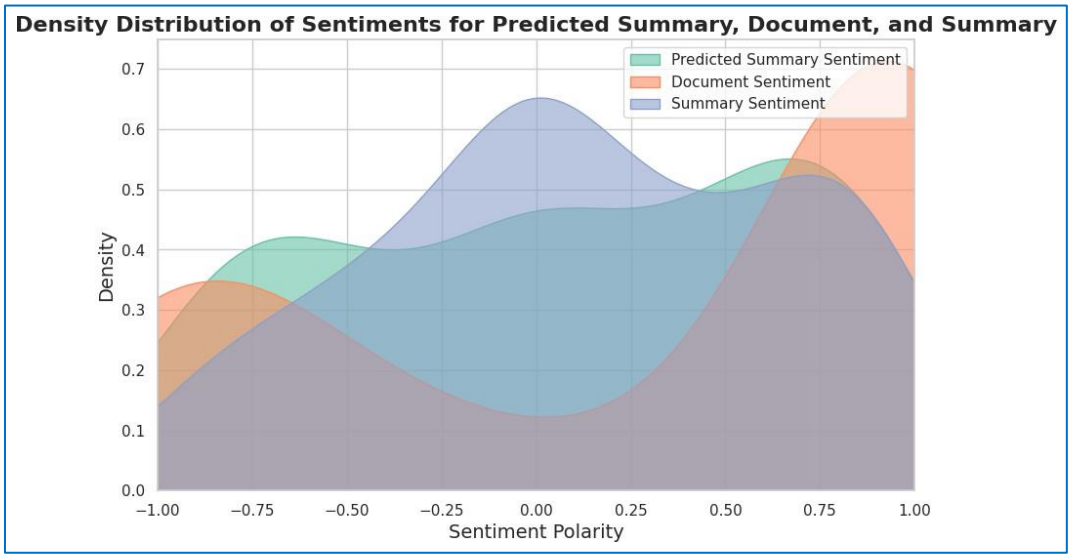
The model's performance was assessed using standard evaluation metrics:

- ROUGE-1: 0.2648
- ROUGE-2: 0.0782
- ROUGE-L: 0.1917
- BLEU Score: 0.0217
- BERT F1 Score: 0.2416

These results highlight moderate performance in capturing key phrases (ROUGE), limited fluency alignment (BLEU), and moderate semantic overlap (BERT F1).

Density Distribution of Sentiments

This plot shows the kernel density estimate (KDE) for the sentiment scores of three different types: Predicted Summary Sentiment, Document Sentiment, and Summary Sentiment. It helps compare the sentiment polarity distribution between these types, where the x-axis represents sentiment polarity, and the y-axis represents the density.

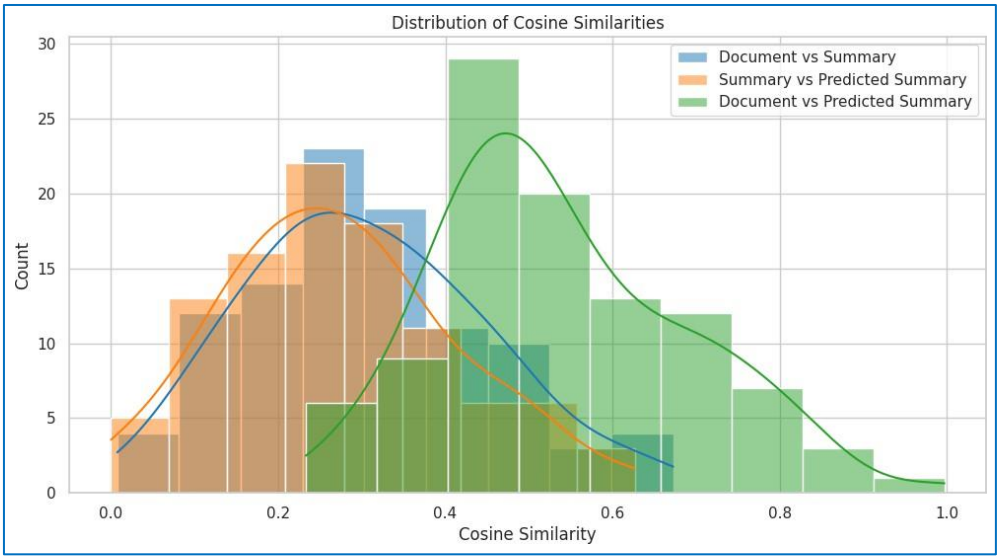


Cosine Similarity Distributions

This histogram displays the distribution of cosine similarities between different text pairs:

- Document vs Summary (Blue)
- Summary vs Predicted Summary (Orange)
- Document vs Predicted Summary (Green)

It visually compares how similar the document, summary, and predicted summary are to each other based on their vector representations.

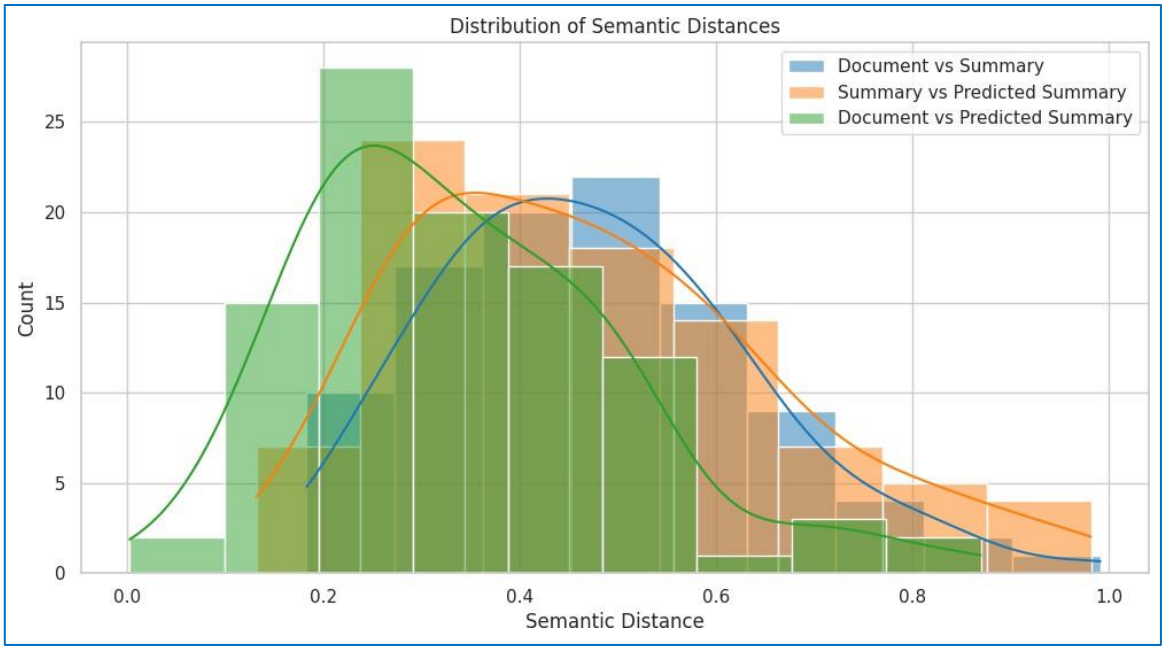


Semantic Distance Distributions

This plot shows the distribution of semantic distances (1 - cosine similarity) between the following pairs of texts:

- Document vs Summary (Blue)
- Summary vs Predicted Summary (Orange)
- Document vs Predicted Summary (Green)

It helps assess the semantic closeness of these text pairs, with lower distances indicating higher similarity.



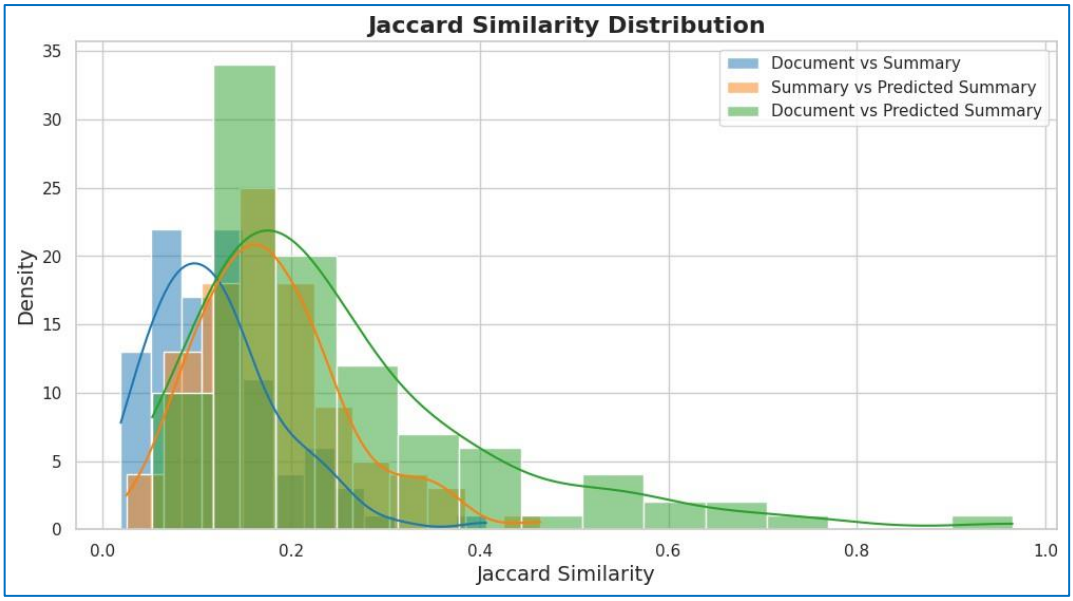
Jaccard Similarity Distribution

This histogram illustrates the distribution of Jaccard similarity between the following pairs of texts:

- Document vs Summary (Blue)
- Summary vs Predicted Summary (Orange)

- Document vs Predicted Summary (Green)

Jaccard similarity measures the proportion of shared words between two texts, and this visualization compares the overlap across different pairs.



UI Design

InsideThread: Reddit Thread Summarization and Hashtag Generation

InsideThread is a Streamlit-based web application designed to transform lengthy Reddit threads into concise summaries, making it easier to extract valuable insights from discussions. It also generates relevant hashtags to help categorize and share the summarized content. The app utilizes advanced Natural Language Processing (NLP) techniques, including text cleaning, slang replacement, and text summarization, powered by the pre-trained BART model.

Key Features

- Text Cleaning and Preprocessing: Removes special characters, URLs, and noisy content from input text to ensure clean data for summarization.
- Slang Replacement: Converts common internet slang terms into their full forms for better understanding and readability.
- Text Summarization: Uses the BART model to generate a concise summary of the input Reddit thread.
- Hashtag Generation: Extracts relevant entities and keywords from the summary and generates hashtags based on them.
- Sharing on Twitter: Allows users to directly share the generated summary on Twitter.

Methodology and Implementation

- Text Preprocessing: Remove Special Characters: The remove_special_characters function uses regular expressions to clean the input text by removing unwanted elements such as HTML tags, URLs, and Reddit usernames.

- Replace Slang: The replace_slangs function replaces common slang words (like "u" for "you" or "smh" for "shaking my head") with their full meanings using a predefined dictionary. This ensures that the text is more formal and easier to summarize.

Text Summarization with BART

The pre-trained BART model is used to generate summaries of Reddit threads. The model is configured to limit the summary length and ensure relevance through parameters like max_length, min_length, num_beams (beam search for better quality), and length_penalty.

The generate_summary function tokenizes the input text and generates a summary based on the specified parameters, ensuring high-quality summaries even for long threads.

Hashtag Generation

- Named Entity Recognition (NER): The generate_hashtags function uses SpaCy to extract named entities (people, organizations, locations) from the summary.
- Keyword Extraction: Noun phrases are extracted from the summary, and stopwords are filtered out. The relevant keywords are then used to generate hashtags, ensuring they are meaningful and relevant to the summary.

Streamlit UI for InsideThread

The Streamlit UI serves as the user interface for interacting with the InsideThread application, allowing users to easily input Reddit threads, generate summaries, and extract hashtags. The UI is designed to be intuitive and user-friendly, providing a seamless experience. Below are the key features of the user interface:

Application Title and Description

The app features a prominent title, "InsideThread", along with a tagline: "Where Data Meets Dialogue", describing the app's purpose.

Instructions

- An expandable section provides users with a step-by-step guide on how to use the application:
- Enter a Reddit thread: Users paste the Reddit conversation into the provided text box.
- Generate Summary: Clicking the "Generate Summary" button creates a concise summary of the thread.
- Generate Hashtags: Clicking the "Generate Hashtags" button generates hashtags based on the summary's key topics.
- Share on Twitter: Users can share the generated summary directly on Twitter.

Input Section

A text area allows users to paste Reddit threads. The input area has a placeholder text to guide users on what to input.

Summary Generation

When the user clicks the "Generate Summary" button, the application cleans the input text by removing unwanted characters and replacing slang. It then generates a summary using the pre-trained BART model. The summary is displayed below the input area for the user to review.

Hashtag Generation

After generating a summary, users can click the "Generate Hashtags" button to receive a list of relevant hashtags based on the summary's content. These hashtags are displayed below the summary.

Sharing Functionality

A "Share on Twitter" button allows users to share the generated summary on Twitter with a pre-filled tweet.

By implementing these features, the InsideThread app provides users with an easy way to summarize Reddit threads and share the key takeaways with the world, while also generating meaningful hashtags for further categorization and discovery.

Conclusion

The Reddit Thread Summarization Project, through its innovative implementation of advanced Natural Language Processing (NLP) techniques, effectively addresses the challenges posed by the vast, unstructured, and nuanced data from Reddit. By leveraging the power of BART, augmented with tools like RAKE and LDA, the project successfully transforms lengthy, complex threads into concise, coherent summaries while preserving context and key themes.

The integration of these methodologies has provided a scalable, efficient, and user-friendly solution, epitomized by the intuitive InsideThread application, which enhances user engagement and promotes informed interaction with Reddit content. This project not only demonstrates technical sophistication but also highlights the potential of data-driven insights to improve digital communication and foster meaningful discourse.

With its comprehensive framework and adaptability, the project sets a strong foundation for future enhancements, such as improved sentiment analysis, live deployment, and seamless integration with other platforms. The work encapsulates a perfect harmony of technological excellence and practical utility, offering a transformative approach to navigating the dynamic world of social media data.