Text Summarization & its Applications

Stephanie Cao, Tanay Chandak, Madelyn Dempsey, Riya Setty

TABLE OF CONTENTS

Introduction

Formal definition, purpose, existing methods, and literature review

What Is Text Summarization?

Formal Definition

Text summarization is **the process of distilling key information from a longer text**, condensing it into a shorter version while maintaining its essence and crucial points. It involves extracting essential content to create a concise summary that provides an overview of the original material, aiding in quick comprehension without the need to read the entire text.

Example Review

This may not be a fair review of the book, as I did not finish. I read perhaps a quarter and finally gave it up, thinking of all the other books I had waiting. It simply did not capture my attention.

Example Gold Standard Summary

Could not finish

1 How much data is generated every minute?

Source: Domo

41,666,667
messages shared by WhatsApp users

1,388,889
video / voice calls made by people worldwide

404,444
hours of video streamed by Netflix users

347,222
stories posted by Instagram users

150,000
messages shared by Facebook users

147,000
photos shared by Facebook users

As text (human-created or generated) increases, so too must our ability to form effective and efficient summaries.

Purpose

Motivations

- **Time Efficiency**
  - Streamlining Information
  - Ex: Professionals can efficiently grasp key points from reports or articles, optimizing time for decision-making.
- **Content Consumption**
  - Enhanced Accessibility
  - Illustration
- **Accessibility to Readers**
  - Inclusivity
  - Impact

Relation to Class

- **Understanding Semantics:** Text summarization involves grasping meaning, aligning with NLP's focus on semantic relationships in language.
- **Information Extraction:** Summarization distills key details, linking to NLP's role in extracting valuable information.
- **Language Generation:** Summarization models create concise summaries, reflecting NLP's emphasis on generating language.

Existing Methods

Extractive Methods: Identify important sentences using algorithms like TextRank or graph-based techniques.

Abstractive Methods: Generate summaries using models like LSTM, Transformer-based (BERT, GPT), or pointer-generator networks.

Hybrid Approaches: Combine extraction and abstraction for more comprehensive summaries.

Key Tools: Libraries like NLTK, Gensim, and frameworks such as Hugging Face's Transformers provide tools and pre-trained models for text summarization tasks.

Literature Reviews

1. The Impact of Local Attention in LSTM for Abstractive Text Summarization
   a. "Amazon Fine Food Reviews" summaries with varying attention evaluated with ROUGE-1 and ROUGE-2
2. Text Summarization Approaches Using Machine Learning & LSTM
   a. Importance of extractive v. abstractive distinction
   b. Approaches to extractive include TextRank, linguistic features, TF-IDF
3. Text Summarization of Articles Using LSTM and Attention-Based LSTM
   a. As people look at long, important articles, they often turn away before finishing
   b. Hyperparameters such as batch size and epochs in LSTM are influential
   c. Bidirectional LSTM not as effective as LSTM for this task

Data

The Data

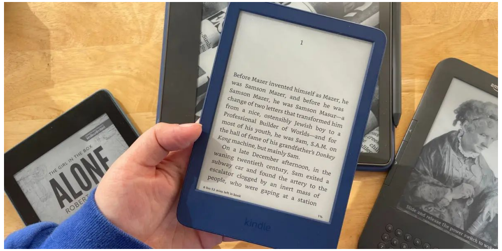
Considerations

- Input document size
- Richness of information
- Summary size
- Labeling (i.e. human generated summaries vs extractive)

Data Sources

- Book reviews (shorter): Amazon reviews: Kindle store category
- Newspaper articles (longer): Cornell Newsrooom

Kindle Reviews Dataset



Overview

Dataset of **product reviews** from Amazon Kindle Store category from May 1996 - July 2014

Size: 982,619 rows

Average length: 109.08 words

Sample Review

I enjoy vintage books and movies so I enjoyed reading this book. The plot was unusual. Don't think killing someone in self-defense but leaving the scene and the body without notifying the police or hitting someone in the jaw to knock them out would wash today.Still it was a good read for me.

Sample Summary

Nice vintage story

Cornell Newsroom Dataset

Overview

Dataset of **articles** and summaries from 38 major publications. The summaries are obtained from search and social metadata between 1998 and 2017 and use a variety of summarization strategies combining extraction and abstraction.

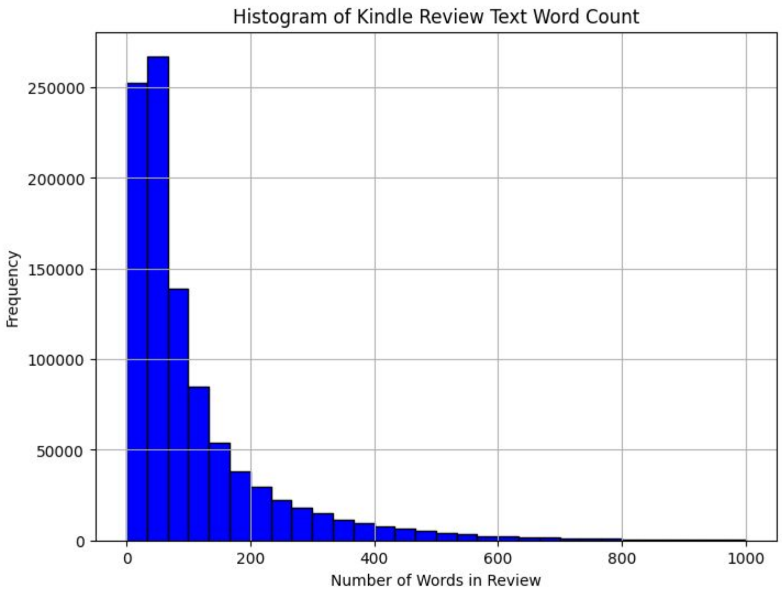
Size: ~1,300,000 rows
Average length: 654.32 words

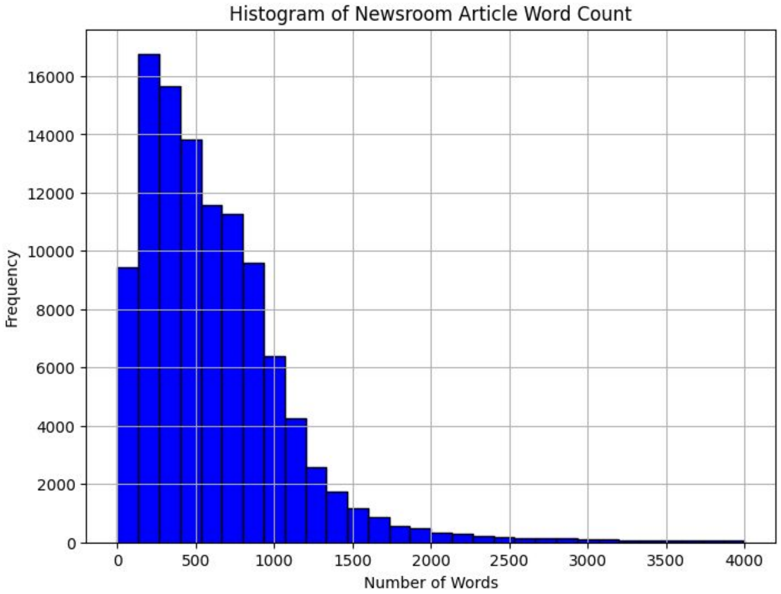
Sample Article

“By MATT SCHWARTZ in Houston and WENDELL JAMIESON in New York Daily News Writers\n\nSaturday, October 14th 1995, 4:22AM\n\nBleeding from a massive chest wound, Tejano star Selena cried, \"Help me! Help me! I've been shot!\" and then named her …”

Sample Summary

"Bleeding from a massive chest wound, Tejano star Selena cried, \"Help me! Help me! I've been shot!\"and then named her killer with her dying breath. Shaken witnesses yesterday told a spellbound Houston courtroom how the blood-covered, mortally wounded 23-year-old Hispanic singing sensation burst into the lobby of the Corpus Christi Days Inn last March 31. Gasping for breath, Selena told motel workers that Yolanda Saldivar the president of her fan club shot"


Histogram of Kindle Review Text Word Count
Frequency
250000
200000
150000
100000
50000
0
0
200
400
600
800
1000
Number of Words in Review


Histogram of Newsroom Article Word Count
Frequency
16000
14000
12000
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
2500
3000
3500
4000
Number of Words

Baselines

Baseline Evaluation: ROUGE-1

Overview

- Focuses on the overlap of n-grams between a model's output text and a reference text
- Measures precision, recall, and f1 based on n-gram overlap

ROUGE-N

$$= \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

Considerations

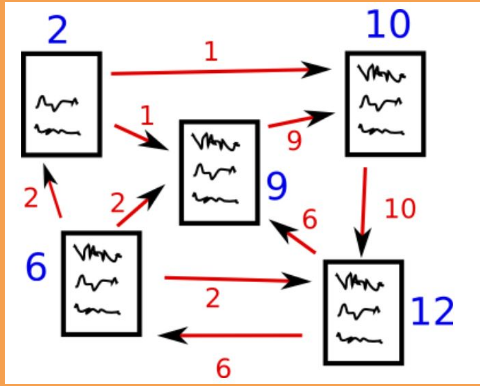
- To what extent do our generated summaries include extraneous information?
- To what extent do our summaries capture the most important information?

Implications

- Best suited for summaries written in prose

Simple Baseline: Extractive

Extractive Summary: extracting the most relevant information from the input
Implementation: TextRank summary using a limit of k sentences and 10 phrases and noun stopwords
Results: 0.0373 rouge score



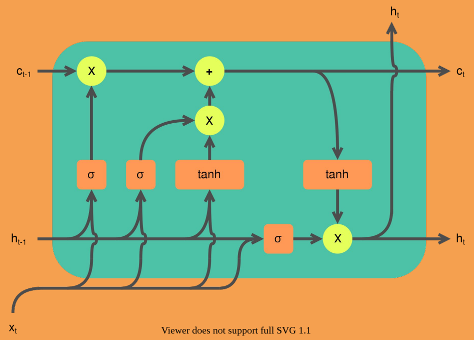
Better Baseline: Abstractive

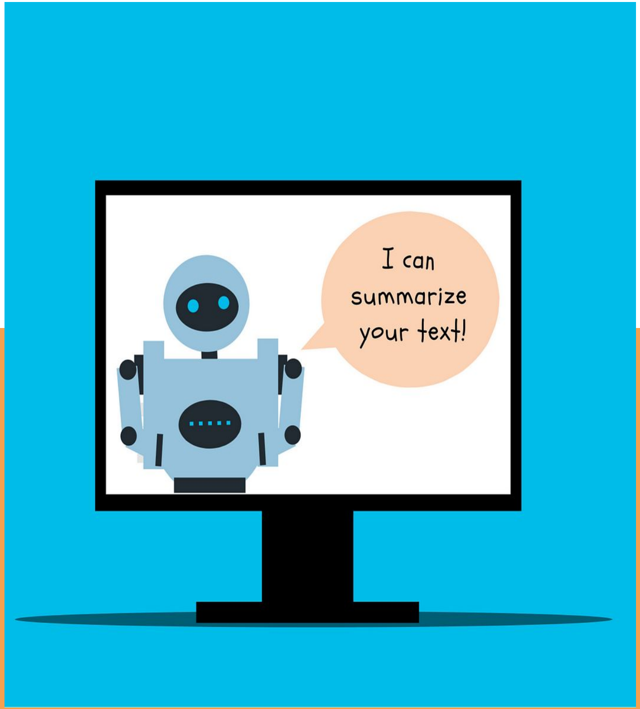
Abstractive Summary: generation through understanding context and meaning.

Implementation: LSTM (built from scratch). Creates new sentences instead of selecting from the original text (unlike extractive summarization).

Results:

- Achieved best results with 100,000 data rows, 50 epochs, and a batch size of 128.
- Lower data volume or epochs resulted in nonsensical or inaccurate summaries.





Improving Our Model

Rouge Score with T5

About T5: Short for Text-to-Text Transfer Transformer, T5 is a versatile machine learning model that converts all natural language processing tasks into a text-to-text framework. It's pretrained on a large text corpus and fine-tuned for specific tasks, which allows it to excel in a variety of NLP challenges with its unified approach.

Our Fine-Tuning Approach: Utilize pretrained model from Hugging Face and fine-tune T5 model on varying number of rows from the Kindle Dataset discussed before. Trained on a smaller summarization set with a GPU in order to provide workable results using an efficient amount of time.

Findings: When compared with the abstractive and extractive models, our fine-tuned T5 model does a lot better, which makes sense due to its more complex architecture and pre-trained nature.

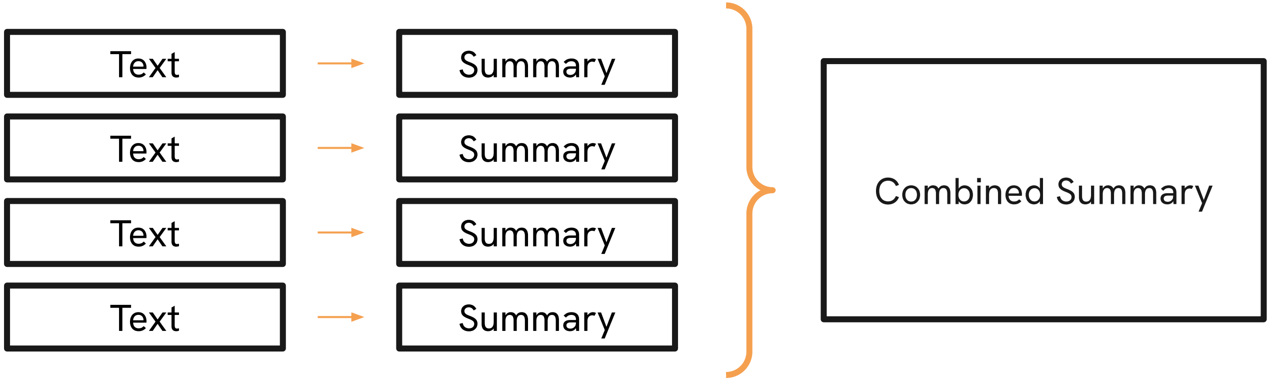
# of Rows	F-Score
1000	0.1142
5000	0.1150
10000	0.1165

# of Epochs	F-Score
3	0.1164
5	0.1152
10	0.1163

Impact of One-Shot vs. Paragraph-Based summarization

Process:

1) Finetune T5 using the shorter Kindle reviews dataset
2) Analyze the performance of the fine-tuned model on the longer Newsroom dataset using two approaches:
   a) Paragraph-stitched summary
   b) Full-text summary
3) Evaluate the summaries with ROUGE-1 evaluation metric



Findings with Precision and Recall

Precision: how much is relevant information
Recall: how much information in gold-standard appears in the generated summary

Conclusions:

- Paragraph-stitched (P) had high recall
- Full-text summaries (F) had higher precision

# of Rows	(P) Precision	(P) Recall	(F) Precision	(F) Recall
1000	0.064623	0.590655	0.230250	0.233150
5000	0.063641	0.597846	0.210920	0.205635
10000	0.063003	0.608440	0.24971	0.247644

# of Epochs	(P) Precision	(P) Recall	(F) Precision	(F) Recall
3	0.064623	0.590655	0.230250	0.233150
5	0.063660	0.602523	0.199578	0.203121
10	0.064298	0.610227	0.225347	0.233420

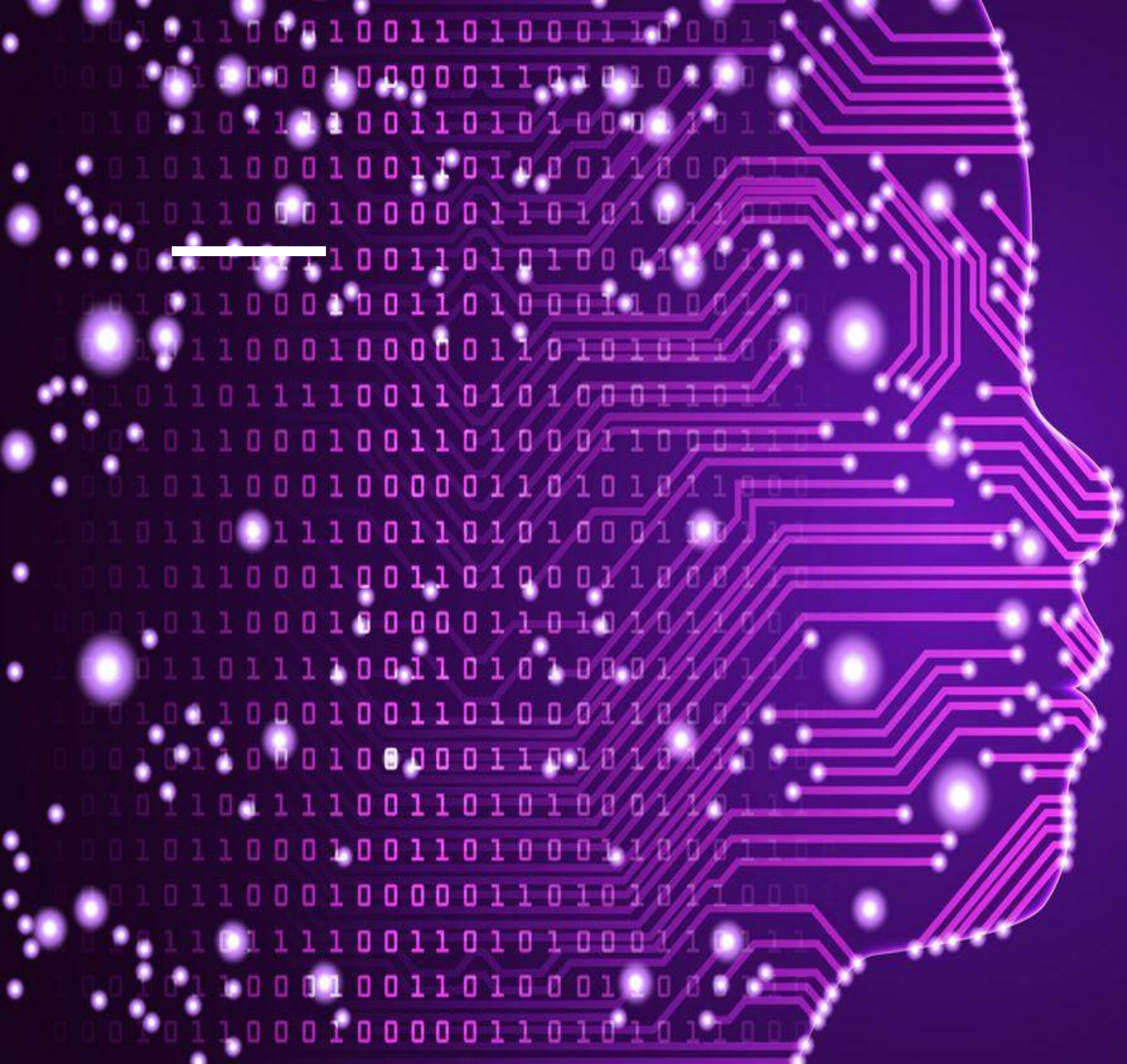
Sample Summary

Input Text: Wallace Robinson, the 6-foot-8-inch center who led the St. Louis University basketball team in rebounds, was dismissed from the team on Jan. 2 because of a fight in which he broke the nose and blackened the eyes of his coach, Ron Ekker, according to a published report. The fight, the report said, occurred in Ekker's hotel room in Indianapolis after Robinson and another player had missed the bus for a game at Butler, which the Billikens won.\n\nEkker's announcement said that Robinson had been dropped from the team for unspecified disciplinary reasons, but the St. Louis Post Dispatch reported the details of the incident. "It's been a terrible time," Ekker said. "I hope it all stops soon because I don't feel I've been able to do my job. This affects me and I'm sure it affects the players. We're trying to overcome it." ... Hugh Johnson, the basketball coach at Richwood High School in West Virginia, has denied charges that he paid starters on his team between 50 cents and $1 for each assist or rebound. The charges were made by a citizens group that has requested a hearing before the country board of education and urged the suspension of the coach.

Gold Standard: Wallace Robinson, the 6-foot-8-inch center who led the St. Louis University basketball team in rebounds, was dismissed from the team on Jan. 2 because of a fight in which he broke the nose and blackened the eyes of his coach, Ron Ekker, according to a published report. The fight, the report said, occurred in Ekker's hotel room in Indianapolis after Robinson and another player had missed the bus for a game at Butler, which the Billikens won.

Paragraph-Stitched Summary: a fight broke the nose and blackened the eyes of his coach, a report said. the fight occurred in Ekker's hotel room in Indianapolis after Robinson missed the bus for a game at Butler. "It's been a terrible time," said Ekker. "I'm sure it affects the players'

Full-Text Summary: 6-foot-8-inch center was dismissed from the team in a fight. he broke the nose and blackened the eyes of his coach, Ron Ekker, according to a report.



Applications: Sentiment Analysis

Applying Sentiment Analysis

Goal: *analyze whether or not our text summaries are able to maintain the sentiment of the original text.*

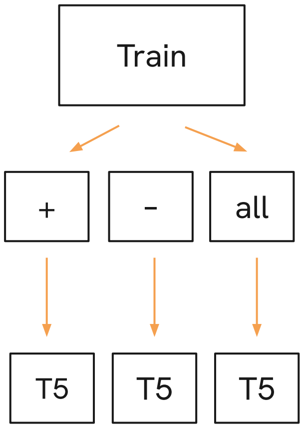
Motivation:

- **Preserving Context and Meaning**
  - Ensuring summaries capture the emotional core is crucial for reader comprehension
- **User Experience and Engagement**
  - Maintaining accurate sentiment in summaries is vital for reader trust and engagement.
- **Ethical Considerations**
  - Preserving original sentiment in summaries is crucial to avoid misrepresentation.
- **Accuracy and faithfulness**
  - Assessing how well summaries convey the original emotional message for accuracy and integrity.

Training Process

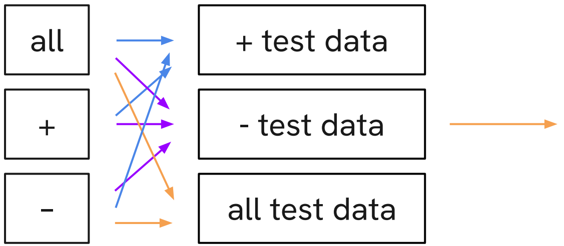
Step 1

Split up our data into test and train dataset and classify sentiment. For the test data. Then fine tune T5 only on Positive / Negative for each classification



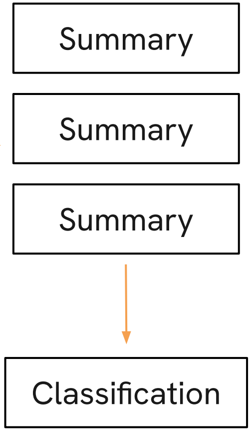
Step 2

From the test data, classify the data into positive, negative, and then generate a summary using each of the T5 models



Step 3

Classify sentiment of each summary using Hugging Face's transformer-based sentiment classification model.



Using Positive and Negative Models

	Full Data	Positive Data	Negative Data
Full Model	71.80 (P) 28.20 (N)	84.92 (P) 15.08 (N)	29.81 (P) 70.19 (N)
Positive Model	73.00 (P) 27.00 (N)	85.71 (P) 14.29 (N)	34.62 (P) 65.38 (N)
Negative Model	66.00 (P) 34.00 (N)	80.42 (P) 19.58 (N)	21.15 (P) 78.85 (N)

(P): % classified as positive
(N): % classified as negative

Analyzing Change in Sentiment with Varied Text Length

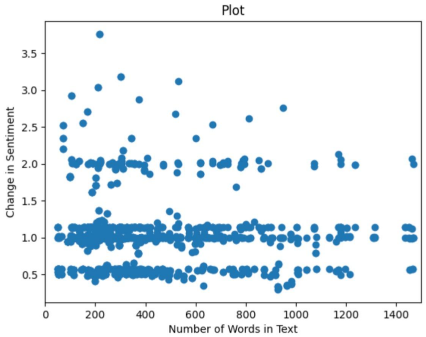
Conclusions:

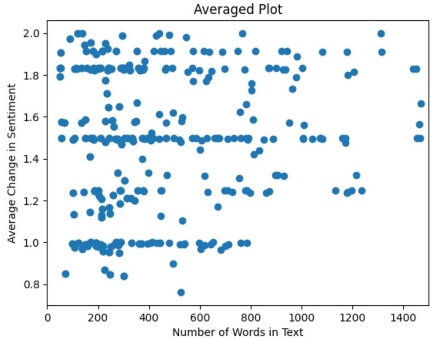
- No clear correlation between length of text and change in sentiment between datasets
- Higher variation in change of sentiment among Newsroom articles

Change in Sentiment vs # Words

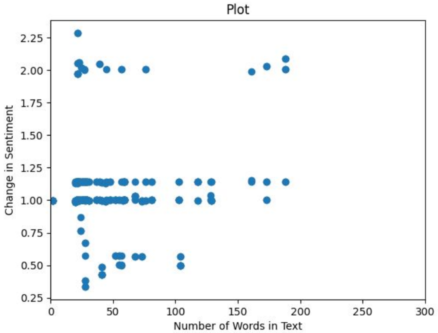
Avg Change in Sentiment vs # Words

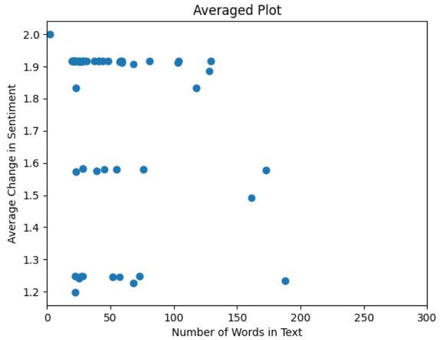
Newsroom Data





Kindle Data





Thank you!