Text Summarization Using NLP

Stephanie Cao and **Tanay Chandak** and **Madelyn Dempsey** and **Riya Setty**
University of Pennsylvania / Philadelphia, PA
caosteph@seas.upenn.edu
tanayc@seas.upenn.edu
dmadelyn@seas.upenn.edu
rsetty@seas.upenn.edu

Abstract

In an era where the volume of written content is expanding globally, the ability to condense information succinctly is becoming increasingly vital. This paper delves into the world of text summarization, a key application for large language models, highlighting the essential building blocks for constructing both small and large-scale computational models. We emphasize the need for understanding and improving the accuracy of summaries produced by smaller models, built from the ground up. Our exploration begins with extractive summarization techniques that utilize the original text, followed by a shift to abstractive models. We provide a comprehensive examination and implementation of these models, including LSTM and the transformer-based T5 model, to assess different summarization strategies. Additionally, we investigate novel aspects of text summarization, such as the impact of summarizing individual paragraphs versus entire texts, and the effect of fine-tuning models on texts with specific sentiments. This paper aims to enhance the understanding of text summarization, exploring its various facets and its importance in managing the ever-growing digital information landscape.

1 Introduction

1.1 The Task

In this paper, we harness the power of Natural Language Processing and apply it to the task of summarizing text. Informally, our task involves taking human-generated input text such as news articles and product reviews, and leveraging computational linguistics to generate a concise summary of the input. Throughout our exploration, we implement various NLP techniques and attempt to understand and model the decision process of determining what information in an input text is most relevant to generating an effective summary.

The primary objective in text summarization is to distill the key points, details, and ideas present in the text, and output a condensed version that is both concise and coherent while retaining the most essential information from the text. The key challenges inherent in this task include the need for semantic understanding, preserving coherence, and overcoming limitations on computational power.

1.2 Illustrative Example

Appendix A depicts an example input and output for the problem at hand. In the illustration, we see that the input text is a review for a book called The Diary of a Nobody. At first glance, the review is quite long, and it would be impractical for those interested in reading this book to read multiple reviews of this length to determine whether or not the book is worth reading. In contrast, the summary is very concise and conveys the main idea of the review that the book itself is "amusing".

1.3 Formal Definition

More formally, text summarization is an NLP problem involving generating a condensed version of a text. Typically, the methodology for approaching such a task uses either an abstraction or extraction approach. Extractive summarization involves selecting and arranging the most important sentences or passages from the input text. This approach maintains the language of the original text and uses methods such as TF-IDF, TextRank, and sentence position to select the most appropriate sentences.

Abstractive summarization requires generating new sentences that may not have been present in the original text. It requires a better understanding of context and semantic analysis.

1.4 Motivations

We chose this task for our term project because as college students we consume an immense amount of textual content including textbook readings, papers, and social media posts. Thus, we have experienced the need for concise and comprehensive sum-

maries to optimize our reading time. Additionally, we were interested in the potential applications of a text summarizer such as increasing accessibility by making long and difficult-to-understand passages more easily readable. It is common for readers to be intimidated by long, dense information. Having access to a text summarization model may help make such information more accessible to people of all reading levels.

Our goal in implementing a text summarizer ultimately is to make information more accessible and enhance a reader's ability to consume the main points of a text efficiently and with greater capacity.

2 Literature Review

The exploration of text summarization techniques has been a focal point in recent research endeavors. Three significant papers—'The Impact of Local Attention in LSTM for Abstractive Text Summarization' by Hanunggul, Suyanto, et al.[1], 'Text Summarization Approaches Using Machine Learning & LSTM' by Sirohi, Bansal, and Rajan[2], and 'Text Summarization of Articles Using LSTM and Attention-Based LSTM' by Kumar, Kumar, Singh, and Paul [3] — shed light on diverse methodologies, challenges, and future pathways in the domain.

Kumar, Kumar, Singh, and Paul's paper 'Text Summarization of Articles Using LSTM and Attention-Based LSTM' underscores the vital role of text summarization, particularly in facilitating access to extensive digital content. Their exploration of extractive and abstractive summarization, supported by experiments with diverse datasets and attention-based LSTM models, emphasizes the effectiveness of LSTM architectures. Notably, the paper discusses the significance of ROUGE metrics in evaluating text summaries, setting the stage for future enhancements in attention layers and dataset scalability.

Hanunggul, Suyanto, et al.'s study focuses on how altering attention mechanisms within an LSTM affects its performance in text summarization. Testing 'global' and 'local' attention types, the researchers used the "Amazon Fine Food Reviews" dataset, cleaning and filtering summaries between 25-300 characters. Evaluating the generated summaries using ROUGE metrics revealed that the LSTM with global attention excelled in ROUGE-1 scores, while local attention of 5 performed best in ROUGE-2. The dataset's casual summaries, rich in colloquial terms not well represented in GloVe embeddings, negatively affected model performance. With an average of 38 tokens per summary and 5-10 tokens per sentence, the larger local attention window proved less successful, potentially due to dataset characteristics. The study suggests exploring syllable-based models for improved performance—a potential extension for future research. Overall, it highlights the efficacy of different attention mechanisms in enhancing model performance in text summarization.

In 'Text Summarization Approaches Using Machine Learning & LSTM,' Sirohi, Bansal, and Rajan delve into extractive and abstractive summarization methods. They outline various techniques encompassing statistical, topic-based, clustering, semantic, and machine learning-driven approaches for both extractive and abstractive summarization. By highlighting the advantages and drawbacks of these methodologies, the paper underscores the importance of effective summaries amidst the burgeoning text landscape.

Collectively, these papers underscore the multifaceted nature of text summarization techniques, from nuanced attention mechanisms and diverse summarization approaches to the challenges posed by informal language and dataset representation. The studies serve as crucial guiding posts for future advancements in refining text summarization models and methodologies, offering insights into potential research directions for enhancing summarization efficacy amidst the evolving text landscape.

3 Experimental Design

3.1 Data

The first dataset we used in our task was the Cornell Newsroom dataset, a large dataset for training and evaluating summarization systems.[4] It contains 1.3 million articles and summaries written by authors and editors in the newsrooms of 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. The dataset has three

[1] https://ieeexplore.ieee.org/abstract/document/9034616

[2] https://revistageintec.net/old/wp-content/uploads/2022/03/2526.pdf

[3] https://link.springer.com/chapter/10.1007/978-981-16-7996-4_10

[4] https://lil.nlp.cornell.edu/newsroom/index.html

large files for training, dev, and test sets - each titled train.jsonl, dev.jsonl, and test.jsonl, respectively.

Each of these three files uses the compressed JSON line format, and an example of the format can be referenced in Appendix B. Each line is an object representing a single article-summary pair. The entire dataset contains 1.3 million available rows for exploration. The number of rows in our dataset split can be seen in Table 1. As described in

Table 1: Newsroom Dataset

Dataset Size	1,321,995 articles
Training Set	995,041 rows
Dev Set	108,837 rows
Test Set	108,862 rows
Mean Article Length	658.6 words
Mean Summary Length	26.7 words
Total Vocab Size	6,925,712 words
Occur 10+ times in text	784,884 words

the paper by Max Grusky, Mor Naaman, and Yoav Artzi, the Newsroom summaries were extracted via metadata available in the HTML pages of articles. [5] The summaries were often written by newsroom editors and journalists via social media crawls. A full histogram displaying the frequencies of words in the Newsroom dataset can be seen in Appendix G for reference. The most commonly used metadata fields for this task were the og:description, twitter:description, and description. Sample inputs and outputs from the test set for this dataset can be seen in Table 2.

The second dataset we used in our experimentation was the Kindle Reviews Dataset containing 982,899 total rows of Kindle book reviews.[6] Each row contains data about the product ID, review helpfulness rating, overall product rating, reviewText, reviewTime, reviewerID, reviewerName, summary, and unixReviewTime. The full histogram displaying word frequencies in the Kindle Dataset can be seen at Appendix F. The data in this dataset was obtained from Amazon Product Reviews, a large crawl of product reviews from Amazon containing 82.83 reviews total. While performing an exploratory analysis on this dataset, we found that approximately 77.71% of the input text from this dataset had a positive sentiment using a Huggingface sentiment analysis pipleline, while 22.29% were negative. When doing the same calculation, we found that approximately 79.4% of

Table 2: Newsroom Dataset Sample Input and Output

Review Text	Summary
"Political reversals at home and continued bad news from Iraq have dragged President Bush's standing with the public to a new low, at the same time that Republican fortunes on Capitol Hill also are deteriorating, according to the latest Washington Post-ABC News poll. The survey found that 38 percent of the public approve of the job Bush is doing, down three percentage points in the past month and his worst showing in Post-ABC polling since he became president. Sixty percent disapprove of his performance. With less than seven months remaining before the midterm elections, Bush's political troubles already appear to be casting a long shadow over them..."	"Political reversals at home and continued bad news from Iraq have dragged President Bush's standing with the public to a new low, at the same time that Republican fortunes on Capitol Hill also are deteriorating, according to the latest Washington Post-ABC News poll."
"Every month, the National Snow and Ice Data Center in Boulder, Colo., puts out a news release about how much ice is floating on the cold seas at the top of the world. Those who follow this obscure bit of news will know that last month marked the lowest extent of Arctic sea ice on record for June, going back to the beginning of satellite observations in the late 1970s. And summer still has a few months to go. Arctic sea ice typically shrinks until mid-September, when darker nights and colder temperatures come, and the ice cover begins to expand again. We don 2019t yet know if the approaching yearly minimum this September will be the historical low, but it seems on track to possibly match, or even exceed, the previous record minimum, which occurred in 2012..."	"The melting away of Arctic ice because of climate change won 2019t raise sea levels, but what it means for the world is one of the big, complicated questions in science right now."

the associated gold-standard summaries were positive while 20.6% were negative. This suggests that there might be a slight bias towards positivity in the gold standard summaries. This is an important consideration to remember as we investigate extensions towards sentiment later in this paper. See Table 3 for sample reviews and gold standard summaries from this dataset.[7]

Table 3: Kindle Dataset Sample Input and Output

Review Text	Summary
"Omg, I loved it! This book is so fun to read especially if you love playing Trivia! Thanks!"	"Awesome Book!"
"My boyfriend and I love Game of Thrones, so we decided to get a hold of this book and I am so happy we did! We have been having so much fun playing trivia. There are really good questions in this book it made the trivia game a lot of fun! Can't wait to have friends over to play."	"Great game!"
"I did sort of enjoy reading this novel. Not as much as I could have if the author had really sat with it and given it a prolonged massage, but I did pour it over my head as it were and it was fine. But I won't read it again."	"It's alright."

[5] https://arxiv.org/pdf/1804.11283v2.pdf

[6] https://www.kaggle.com/datasets/bharadwaj6/kindle-reviews/data

[7] https://cseweb.ucsd.edu/~jmcauley/datasets.html

3.2 Evaluation Metric

For our evaluation measure, we decided to use ROUGE-N in which we calculate precision and recall based on the number of overlapping n-grams between a given system input and gold standard text. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation and comes in several different flavors including ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. We chose this evaluation metric because it allows us to assess the quality of our system's summaries against the dataset's summaries. It determines the quality of the summary by f1 score and therefore tells us how similar the system's output was to the gold standard summary by helping to capture the most important content. Using n-grams in ROUGE-N was essential for our task because we needed to consider how we would want to evaluate how both the strong and simple baselines performed using the same metric. Since we are using both abstractive and extractive techniques, it made sense to use a metric that involved small units of the output such as n-grams, since we may have summaries that simply extract exact phrases from the text or system-generated summaries. The best way to leverage both systems was to compare the n-gram similarities of the system output and gold standard summaries.

In practice, we chose ROUGE-1 as our final evaluation score since it made the best sense given that the Kindle gold standard reviews are quite short. We used the following equation to calculate the ROUGE scores:

$$\text{ROUGE-N} = \frac{\sum_{S \in ReferenceSummaries} \sum_{gram_m \in S} Count_match(gram_n)}{\sum_{S \in ReferenceSummaries} \sum_{gram_m \in S} Count(gram_n)}$$

The paper explaining this equation and its use as a metric for text summarization can be found here. Past publications using the same metric for text summarization tasks include ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks, A Critical Analysis of Metrics Used for Measuring Progress in Artificial Intelligence, and Text Summarization Using Extractive Techniques.

3.3 Simple Baseline

In our simple baseline, we implemented an extractive summarization technique. Given a piece of text for a training sequence, we compute the text rank for each sentence in the input sequence. Then given an input n, we choose the n sentences of highest rank and output these sentences as the summary. The TextRank algorithm is a graph-based ranking algorithm that uses cosine similarity between the vector representation of sentences to establish connections and then uses the PageRank algorithm on the sentence graph to determine the importance of each sentence.

The key idea is that sentences containing important information are likely to be connected to other important sentences. By iteratively computing scores based on these connections, TextRank identifies the most important sentences for summarization. TextRank leverages cosine similarity to establish connections between sentences, and then the PageRank algorithm ranks the importance of these sentences. Finally, we return the n most important sentences as the output summary.

An important factor in tuning this simple baseline is the determination for n, the number of sentences to extract from the input text to generate the output summary. Sample outputs from our extractive baseline when run on the Kindle Review test dataset with n = 3 can be seen in Table 4. The overall Rouge score for the Kindle Dataset run on this simple baseline was 0.0373. Note here that our gold standard summary is much shorter than the output summaries from our model given that n = 1 in the ROUGE calculation. Thus, it is understandable to expect a low score with this input. Initially, our goal was to run the simple baseline on the Newsroom dataset only; however, we quickly encountered difficulties with the Newsroom dataset while running our better baseline since the sheer volume of the input text was overwhelming for our LSTM algorithm. Looking at the summaries themselves, we see that our model was indeed able to capture the most relevant pieces of text. In the next section, we will investigate how our model performs using a more difficult abstractive approach.

4 Experimental Results

4.1 Published Baseline

For the strong baseline, we took the route of abstractive summarization; rather than selecting sentences from the original text like extractive summarization, abstractive summarization generates new sentences with new words. We implemented a sequence-to-sequence architecture with attention-based LSTM layers for the encoding and decoding processes.

A huge issue we came across was training the

Table 4: Simple Baseline Sample Output

Review Text	Model Output	Gold Standard Summary
"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."	"I enjoy vintage books and movies so I enjoyed reading this book."	"Nice vintage story."
"This book is a reissue of an old one; the author was born in 1910. It's of the era of, say, Nero Wolfe. The introduction was quite interesting, explaining who the author was and why he's been forgotten; I'd never heard of him. The language is a little dated at times, like calling a gun a #34;heater.#34; I also made good use of my Fire's dictionary to look up words like #34;deshabille#34; and #34;Canarsie.#34; Still, it was well worth a look-see."	"The language is a little dated at times, like calling a gun a #34;heater.#34; I also made good use of my Fire's dictionary to look up words like #34;deshabille#34; and #34;Canarsie.#34; Still, it was well worth a look-see."	"Oldie"
"A beautiful in-depth character description makes it like a fast pacing movie. It is a pity Mr Merwin did not write 30 instead only 3 of the Amy Brewster mysteries."	"It is a pity Mr Merwin did not write 30 instead only 3 of the Amy Brewster mysteries."	"Nice old fashioned story"

model in a reasonable amount of time on the news dataset, due to the sheer amount of text that had to be processed with our limited computing resources. To temporarily combat this problem, we worked with a different dataset of reviews from the Amazon Kindle store, which was both smaller in overall size and consisted of smaller data points of text to process. However, even despite this, our model performed worse than expected due to our inability to efficiently train and the general shortage of data. Moreover, the model complexity didn't seem to address all the nuances of general text summarization, and without any pre-trained weights, it wasn't able to produce strong examples.

4.2 Extensions

For our first extension, we decided to use T5 to improve our abstractive text summarization. We were interested in seeing what it would look like to summarize each individual paragraph in an article and stitch those summaries together as opposed to just summarizing the entire article, so we took the approach of fine-tuning T5 on our Kindle dataset (in which each text is a paragraph), and testing it on the Newsroom dataset (in which each text consists of multiple paragraphs), then comparing the two methods.

Below are tables outlining precision and recall scores with different hyperparameters; (P) represents the stitched paragraph summary, and (F) represents the full-text summary.

Table 5: Hyperparameter Tuning - Number of Rows

# of Rows	(P) Precision	(P) Recall	(F) Precision	(F) Recall
1000	0.064623	0.590655	0.23025	0.23315
5000	0.063641	0.597846	0.21092	0.205635
10000	0.063003	0.60844	0.24971	0.247644

Table 6: Hyperparameter Tuning - Number of Epochs

# of Epochs	(P) Precision	(P) Recall	(F) Precision	(F) Recall
3	0.064623	0.590655	0.23025	0.23315
5	0.06366	0.602523	0.199578	0.203121
10	0.064298	0.610227	0.225347	0.23342

The full-text summaries seemed to consistently yield higher precision scores, while the paragraph-stitched summaries consistently had significantly higher recall scores. Fine-tuning the model with more rows or epochs generally seems to result in a higher recall, whereas the precision either changes a minuscule amount or remains the same. Example outputs for this extension on two example texts from the Newsroom Test set can be found in Appendix D. Example 1, Appendix D presents a text that had the highest precision via ROUGE-1, meaning that it contained the most relevant information from the news story. The words in the Full-Text summary all appear in the Gold Standard summary and are therefore 'relevant,' prompting a precision of 1.0. However, the Paragraph-Stitched Summary contains the phrase "said Ekker," which doesn't appear in the Gold Standard summary, resulting in a lower precision. Both generated summaries are fairly short compared to the Gold Standard, which explains the lower recall scores.

In Example 2, Appendix D, we see an example with high paragraph-recall, so each paragraph of the summary had the most direct similarity to the gold standard using our paragraph-stitched approach. Finally, in Example 3 we see an instance of a summary that had high recall in the full-text approach.

Our simple baseline returned an average F1 score of 0.0373 using the ROUGE-1 metric, whereas the most successful version of this extension returned an F1 score of 0.24867 via the ROUGE-1 metric; this means that the abstractive approach with T5 performed almost 7 times better than our extractive approach with TextRank.

For our second extension, we decided to explore the impact of sentiment in an initial text on its

corresponding summary. We were curious to see how summaries and their sentiment scores would change depending on differently fine-tuned models (specifically with positive text, negative text, and the full dataset). Since this milestone also uses T5, it performs better than our baselines, and we also noticed some interesting patterns in the sentiment classification in the table below. Example output for input summaries from the Newsroom Test set to this extension can be found in Appendix E.

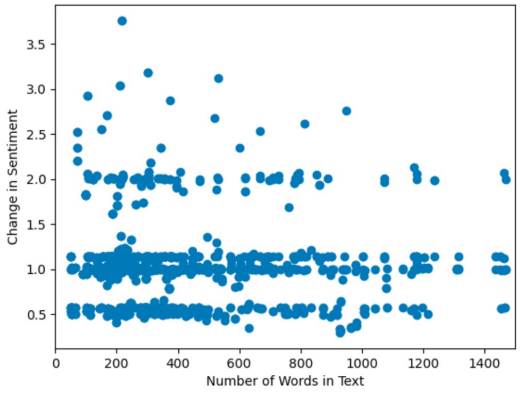
Table 7: Percent of Data Classified by Sentiment

Model	Full Data	Positive Data	Negative Data
Full	71.8 (P) 28.2 (N)	84.92 (P) 15.08 (N)	29.81 (P) 70.19 (N)
Positive	73 (P) 27 (N)	85.71 (P) 14.29 (N)	34.62 (P) 65.38 (N)
Negative	66 (P) 34 (N)	80.42 (P) 19.58 (N)	21.15 (P) 78.85 (N)

Ideally, the summaries generated from the positive model and positive data should all have positive sentiment labels, but our results show that only 85% do. However, our models do seem to perform as expected more than 50% of the time.

We also thought to look at the relationship between the number of words in the initial text and the change and sentiment throughout summaries generated by the different models, but we saw no correlation, as seen in the figure below.

Number of Words in Input Text vs. Change in Sentiment:



4.3 Error Analysis

Our best-performing system was our fine-tuned T5 model from Milestone 3; however, we came across cases where it performed worse than our simple baseline.

Take the following review: "This was an enjoyable story that didn't take too long to read. If you're looking for some light reading and warm feelings, this is the book." The gold-standard summary for this is "nice light reading" Our simple baseline produced "if you're looking for some light reading and warm feelings, this is the book," while our fine-tuned T5 produced "a good read. a good read. a good read. a good read. a good read. a good read. a good read. a good read." Here, our T5 summary is repetitive to the point of being nonsensical; and although the summary our simple baseline produced is still long, the extractive approach produced a sentence that at least makes sense and is similar to the gold standard. See Appendix E for a more in-depth sample output from this extension.

We noticed that the reviews are grouped by the book they're referring to. We initially made the mistake of not randomizing our train set selection, so our model was fine-tuned on reviews of a specific genre, hurting our performance. Similarly, we found that as we increased the amount of data we fine-tuned with, the quality of our summaries increased; due to computational limitations, the largest amount of data we ever worked with was 10,000 rows. Thus, we'd assume that much of our error is the result of training on minimal data.

5 Conclusions

Throughout our exploration, the utilization of diverse techniques—such as TextRank for extraction and LSTM-based architectures for abstraction—gave us interesting insights into the realm of text summarization. Notably, the fine-tuning of the T5 model stood out as a superior approach, showcasing promising results. However, our analysis unveiled instances where even this advanced model faced challenges, indicating uncharted territory for further research. Understanding the nuances behind these limitations presents an exciting opportunity to refine and optimize summarization models for varied text types and contexts.

Our investigation into sentiment preservation revealed that, while a majority of texts exhibited preserved sentiment (at an 85% rate), there is still a need for higher alignment, especially in applications like review summaries. Enhancing sentiment preservation mechanisms in summarization remains a critical focus area for achieving more nuanced and contextually accurate summaries.

In the future, there are many research possibilities. Finding ways to make the T5 model work better in different situations and creating strong methods to keep the feelings in summaries aligned are just some of the things that need exploring. Our discoveries show that everyone in the NLP community has a role in making text summarization better

by coming up with new ideas and working together. Better ways of summarizing text can make a big difference in journalism, education, and analyzing data. By staying dedicated to new ideas and working together, we're making text summarization better and more useful for the future.

6 Acknowledgements

1. Mark Yatskar
2. The CIS 5300 TAs for their support
3. Public NLP articles and research made available

References

[1] Blagec, Kathrin et al. "A critical analysis of metrics used for measuring progress in artificial intelligence." ArXiv abs/2008.02577 (2020): n. pag.

[2] Deokar, Varun, and Kanishk Shah. (2021). *Automated Text Summarization of News Articles*. International Research Journal of Engineering and Technology 8.9: 1-13.

[3] Etemad, Abdul Ghafoor, Ali Imam Abidi, and Megha Chhabra. (2021). *Fine-Tuned T5 for Abstractive Summarization*. International Journal of Performability Engineering 17.10.

[4] Ganesan, Kavita. "ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks." arXiv:1803.01937 [cs.IR], version 1, 5 Mar. 2018, https://doi.org/10.48550/arXiv.1803.01937.

[5] Grusky, Max, Mor Naaman, and Yoav Artzi. (2018). *NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 708-719. http://aclweb.org/anthology/N18-1065

[6] Grusky, Max, et al. (2020). *NEWSROOM: A Dataset of 1.3 Million Summaries with Diverse Extractive Strategies*. arXiv. https://ar5iv.org/pdf/1804.11283

[7] Hanunggul, Puruso Muhammad, and Suyanto Suyanto. (2019). *The Impact of Local Attention in LSTM for Abstractive Text Summarization*. 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2019, pp. 54-57. https://doi.org/10.1109/ISRITI48646.2019.9034616

[8] Kumar, H., Kumar, G., Singh, S., Paul, S. (2022). *Text Summarization of Articles Using LSTM and Attention-Based LSTM*. In: Chen, J.IZ., Wang, H., Du, KL., Suma, V. (eds) Machine Learning and Autonomous Systems. Smart Innovation, Systems and Technologies, vol 269. Springer, Singapore. https://doi.org/10.1007/978-981-16-7996-4_10

[9] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004. https://aclanthology.org/W04-1013.pdf

[10] Puspitaningrum, Diyah. (2022). *A survey of recent abstract summarization techniques*. Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 4. Springer Singapore.

[11] Siddiqui, H., et al. "Text Summarization using Extractive Techniques." 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 28-31. IEEE, doi: 10.1109/ICAC3N53548.2021.9725501.

[12] Sirohi, Neeraj & Rajshree, Mamta & Rajan, Siddhi. (2021). *Text Summarization Approaches Using Machine Learning & LSTM*. Revista Gestão Inovação e Tecnologias. 11. 5010-5026. https://doi.org/10.47059/revistageintec.v11i4.2526

[13] Vaswani, Ashish, et al. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems 31: 6000-6010.

Appendices

A Illustrative Example

Review: Thirty years before Sinclair Lewis publishedBabbittand set the standard for smug, self-important middle-class conformity, there was The Diary of a Nobody and Charles Pooter. Pooter, a senior bank clerk in the City renting a home in a London suburb of Holloway, encapsulates Victorian respectability, snobbery, and pretensions. Pooter nearly invariably gets the short end of the stick in his interactions with his two neighbors, Cummings and Gowings; his spendthrift, reckless son Lupin; and the various tradesmen and servants he attempts to bully. Slavishly devoted to his employer, Mr. Perkupp, Pooter tries without much luck to cut his son into the same mold. Instead, Lupin slacks at work and spends his nights engaged in amateur theatrics or carousing with his chums till all hours. What's a father to do? First serialized in Punch in 1888 and 1889, The Diary of a Nobody was published in book form in 1892 and hasn't been out of print since. If you give this slim volume a chance (available for free in the Kindle format), you'll

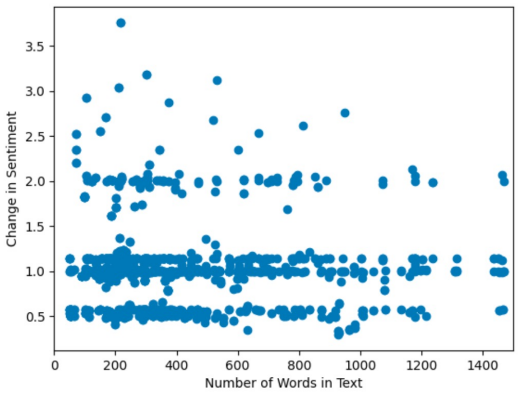
see why. Despite being a century old, The Diary of a Nobody remains quite amusing and is laugh-out-loud funny in parts – particularly in sections dealing with tradesmen or with Lupin's impetuous business dealings or love affairs. As long as there are self-satisfied petit bourgeois snobs, The Diary of a Nobody will continue to entertain.

Gold Standard Summary: a century later, still quite amusing

B Newsroom Data Format


```
{
          "text": "...",
       "summary": "...",
         "title": "...",
       "archive": "http://...",
          "date": 20160302060024,
       "density": 1.25,
      "coverage": 0.75,
   "compression": 12.5,
"compression_bin": "medium",
  "coverage_bin": "low",
   "density_bin": "abstractive"
}
```


C Number of Words in Input Text vs. Change in Sentiment



D Sample Output From Extension 1 Models

EXAMPLE 1

Review 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. Ekker'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 Summary:
"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."

EXAMPLE 2

Review Text:
"The baby came his way—even if Fetty Wap long denied paternity. The '678' star welcomed a baby girl late Tuesday night with former flame Masika Kalysha, TMZ reported. While the 24-year-old has yet to share the news, Kalysha took to Twitter to

share love for her bundle of joy. 'She's here. I've never seen anything so perfect in my entire life. I'm so in love with you #KBM,' the 30-year-old wrote. 'I can't believe she's ours. Best day of my life @fettywap #KBM,' she added. The three-letter hash tag stands for Khari Barbie Maxwell, who was born in a Los Angeles hospital weighing nearly 7 pounds, according to the gossip site. The new mom explained on Instagram that Khari means joyful and queenly in Swahili, which is appropriate for the 'Trap Queen' hit-maker. The rapper, real name Willie Maxwell, reportedly rushed from New Jersey to L.A. to welcome his new daughter. PETITION CALLS FOR FETTY WAP TO PERFORM 'TRAP QUEEN' AT NANCY REAGAN'S FUNERAL When Kalysha, known for 'Love & Hip Hop: Hollywood,' announced the baby news last year, Wap vehemently denied he was the father. 'Congratulations to her and whoever is the (father),' Wap, who has two other children, said in a January interview. The Jersey native is clearly singing a different tune now—months after Kalysha alleged she repeatedly asked him to take a DNA test to prove her claim. The pair met when they collaborated on track the 'Andale' last year. Wap has a 4-year-old son, Aydin, and a 1-year-old daughter, Zaviera, from past relationships."

Gold Standard Summary:
"The baby came his way — even if Fetty Wap long denied paternity."

Paragraph-Stitched Summary:
"The baby came his way...even if Fetty Wap long denied paternity. The '678' star welcomed a baby girl with former flame Masika Kalysha, TMZ reported. The baby girl was born with a baby girl named '678' the 24-year-old has yet to share the news. She took to Twitter to share the love of her bundle of joy. The 24-year-old has yet to share the news. I'm so in love with you #KBM. I've never seen anything so perfect in my entire life. I'm so in love with you #KBM. 'I can't believe she's ours,' she says. Best day of my life @fettywap #kwap #kmswap. The three-letter hash tag stands for Khari Barbie Maxwell, who was born in a Los Angeles hospital weighing nearly 7 pounds. He weighs nearly 7 pounds, according to the gossip site. New mom explained on Instagram that Khari means joyful and queenly in Swahili. The hit-maker is appropriate for the 'Trap Queen' hit-maker. He's a hit-maker. Willie Maxwell, real name Willie, reportedly rushed from New Jersey to L.A. to welcome new daughter. Rushed from New Jersey to L.A. PETITION CALLS FOR FETTY WAP TO PERFORM 'TRAP QUEEN' AT NANCY REAGAN'S FUNERAL. Wap vehemently denied he was the father. Vehemently denied he was the father's father. Vehemently denied the baby news. 'Congratulations to her and whoever is the (father),' she said in a January interview. She has two other children, and she has two other children. Kalysha alleged she repeatedly asked him to take a DNA test to prove her claim. He is now singing a different song. He is clearly singing a different tune now. The pair met when they collaborated on track. They collaborated on track the 'Andale' last year. They met when they collaborated on track the 'Andale.' 4-year-old son, Aydin, and 1-year-old daughter, Zaviera. A 1-year-old daughter, Zaviera, has a 4-year-old son."

Full-Text Summary:
"'678' star welcomes a baby girl. 'I can't believe she's ours,' says a 30-year-old. 'I'm so so in love with you'"

EXAMPLE 3

Review Text:
"U.S. President Barack Obama on Tuesday nominated former Bank of Hawaii chief executive Allan Landon to join the U.S. Federal Reserve's board of governors. Lawmakers had been urging the White House to appoint someone with community bank experience to the board. If confirmed by the U.S. Senate, Landon would have a permanent vote on monetary policy and help implement the 2010 Dodd-Frank financial oversight law. Landon, a partner with private investment fund Community BanCapital, served as chief executive officer of the Bank of Hawaii from 2004 until 2010. BanCapital, based in Portland, Oregon, invests in the debt of community banks. According to the fund's website, Landon, 66, was responsible for "sourcing, evaluating and monitoring investments." He worked as chief financial officer at Bank of Hawaii for four years before becoming CEO. Previously, he had been the CFO at First American in Tennessee and before that had worked at accounting firm Ernst & Young. The nomination

comes as the Fed prepares for its first interest rate increase since 2006, which is expected around mid-year."

Gold Standard Summary:
"If confirmed, Allan Landon would have a permanent vote on monetary policy and help implement the 2010 Dodd-Frank financial oversight law."

Paragraph-Stitched Summary:
"U.S. President Barack Obama nominated former Bank of Hawaii chief executive Allan Landon to join the U.S. Federal Reserve. he also nominated former chief executive of the Federal Reserve. Landon would have a permanent vote on monetary policy. if confirmed, landon would have a permanent vote on the 2010 Dodd-Frank financial oversight law. Landon, a partner with the private investment fund, served as chief executive officer of the bank of Hawaii from 2004 until 2010. he served as chief executive officer of the Bank of Hawaii. BanCapital, based in Portland, Oregon, invests in the debt of community banks. the fund's website was responsible for sourcing, evaluating and monitoring investments. Previously, he had been the CFO at First American in Tennessee. Previously, he had been the chief financial officer at Bank of Hawaii for four years before becoming CEO. the Fed prepares for its first interest rate increase since 2006. the Fed prepares for its first interest rate hike in mid-year-earning. the Fed prepares for its first interest rate increase since 2006."

Full-Text Summary:
"U.S. president Barack Obama nominated former Bank of Hawaii chief executive Allan Landon to join the board of governors. Landon would have a permanent vote on monetary policy and help implement the 2010 Dodd-Frank financial oversight law."

E Sample Output From Extension 2 Models

EXAMPLE 1

Review Text:
"This is only the second book I've read from Robin L. Rotham, but it won't be the last! If you like a little taste of everything in your reads, you will be pleasantly stuffed by the end. The four main characters were interesting in getting to learn some strengths and weaknesses, struggles and rewards. I am starting to read Carnal Compromise right now!"

Gold Standard Summary:
"A FULL Buffet of HEAT!!!"

Positive Model:
Carnal Compromise is the second book I've read from Robin L. Rotham - but it won't be the last. The main characters were interesting in getting to learn some strengths and weaknesses, struggles and rewards.

Negative Model:
I'm starting to read Carnal Compromise right now. I'm starting to learn some strengths and weaknesses, struggles and rewards. I'm starting to read Carnal Compromise right now.

Full Model:
Carnal Compromise is the second book I've read from Robin L. Rotham. The four main characters were interesting in getting to learn some strengths and weaknesses, struggles and rewards.

EXAMPLE 2

Review Text:
"Very different story. No billionaires, etc. Down to earth farmers, hot sex, deep feelings and an interesting tale. Very gritty look at sex and romance in the farming community. Brent and Joe show up at Hake and Mandy's. Mandy is a kinky gal who writes erotic stories. Hake is recovering from a farm injury. Brent and Joe help out by staging a sexy role-play game. Wow, some of this I didn't see coming!! Well done, intense, and fun, too."

Gold Standard Summary:
"Different !!!!!!!"

Positive Model:
"Hake and Mandy is a kinky gal who writes erotic stories. A gritty look at sex and romance in the farming community. A sexy role-play game."

Negative Model:
"A kinky gal who writes erotic stories. A sexy role-play game for the farming community. A

kinky gal who writes erotic stories."

Full Model:
"Hot sex, deep feelings and an interesting story. Hake and Mandy is a kinky gal who writes erotic stories."

EXAMPLE 3

Review Text:
"I really enjoyed this book. I wasn't sure how it would all fall into place when I purchased it, but when I read it I couldn't stop reading till it was done. Then I was sad because it was over. The next one is even better!!"

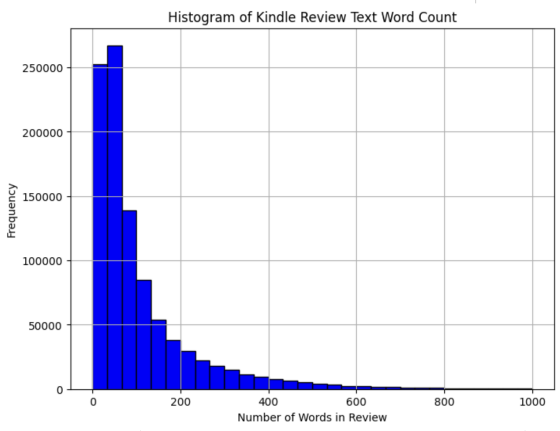
Gold Standard Summary:
"Very good"

Positive Model:
"I enjoyed this book - but it was over....."

Negative Model:
"I really enjoyed this book. I couldn't stop reading till it was done. The next one is even better. The next one is even better. The next one is even better."

Full Model:
"I really enjoyed this book. I couldn't stop reading it till it was done. I was sad because it was over....."

F Word Counts in Kindle Dataset



G Word Counts in Newsroom Dataset