Summarization of Amazon Fine Food Reviews using Deep Learning

Saiem Irfan, Abdulrahman Abdulrazzag, Danyal Ilyas, Burhan Mirza

June 8, 2024

Abstract

This project proposes the development of a neural network model to summarize customer reviews from the Amazon Fine Food dataset. Our objective is to develop a deep learning-based model that not only generates concise summaries reflective of complex consumer sentiment, but also adheres to privacy and ethical considerations by employing differential privacy techniques. The model will be trained on a dataset featuring over half a million reviews, attributed to a quarter million users and covering more than 75 thousand products, ensuring a rich diversity in user opinions and linguistic expressions. We propose a novel sequence-to-sequence neural network architecture with an attention mechanism, specifically adapted to handle the nuances of user-generated content, varying sentence structures, and informal language. The architecture is designed to respect the anonymity of the reviewers and to mitigate bias by preventing preconceived judgments based on reviewer identity. By committing to ethical machine learning practices and continuous stakeholder engagement, we aim to create a robust summarization tool that aids in decision-making without compromising individual narrative integrity or consumer privacy.

1 Introduction

In the digital age, the sheer volume of customer reviews becomes a double-edged sword: rich in data yet daunting to comprehend. Our project is about harnessing deep learning to create succinct, intelligible summaries from the extensive array of reviews on the Amazon Fine Food dataset. This endeavor is not only crucial for assisting consumers in navigating the plethora of opinions but also pivotal for manufacturers who wish to gain an insightful understanding of the public perception of their products. The importance of this task is underscored by the need for quick and informed decision-making in the fast-paced online marketplace, where time and clarity are of the essence.

The novelty of our approach lies in integrating the principles of differential privacy to protect the identities of reviewers, thereby maintaining their anonymity and preventing the influence of biases in the interpretation of reviews. This is meaningful because it ensures that the insights derived from our model are based on content rather than the potentially biased perception of the reviewer's identity. By leveraging cutting-edge advancements in neural

networks and natural language processing, we contend that deep learning is not merely a suitable tool but an innovative avenue for tackling the complexities of summarization in a privacy-conscious manner. This project, therefore, serves as a significant step towards ethical AI, where utility and user confidentiality coexist seamlessly.

2 Background and Related Work

The domain of natural language processing (NLP) encompasses a variety of techniques aimed at understanding and generating human language. One area of NLP that has gained considerable attention is text summarization, which aims to condense lengthy documents into concise summaries without losing critical information.

Early text summarization techniques were largely extractive, identifying key sentences or phrases to stitch together as a summary [5]. However, these methods often resulted in disjointed summaries that lacked the cohesiveness of human-generated text. Deep learning facilitated the use of abstractive summarization, where neural networks emulate human-like paraphrasing of texts' main points [4].

The work of Sutskever et al. [5] on sequence-to-sequence (seq2seq) learning laid the foundation for modern abstractive summarization. It introduced a framework where sequences of text are encoded into a fixed-dimensional context vector from which the target sequence is generated. Building on this, Bahdanau et al. [2] introduced an attention mechanism, enabling the model to dynamically focus on different parts of the source text during the generation of each word in the summary. This innovation was crucial in overcoming the limitations of the seq2seq model, particularly for longer sequences where important information could be lost.

Our project is inspired by these developments and seeks to apply them to the domain of user-generated content, which poses unique challenges. User reviews are characterized by diverse linguistic structures, informal language, and often contain subjective sentiments and slang that standard models may not capture effectively. Additionally, reviews can vary significantly in length and quality, requiring a model that can discern and synthesize key points without being misled by noise or irrelevant details.

In response to these challenges, we propose a novel adaptation of the seq2seq framework, utilizing the T5-base model as our starting point. The T5 model, introduced by Raffel et al. [3], embodies a text-to-text transfer learning approach where a variety of NLP tasks are reframed as text generation problems. Its unified structure and pre-training on a diverse corpus make it a compelling choice for our summarization task. Unlike traditional NLP models, T5 is pre-trained on a multitask mixture, allowing it to leverage learned representations from related tasks, which is particularly beneficial when dealing with the idiosyncrasies of user-generated text.

3 Data

Our model is being developed on a diverse dataset from Amazon's fine food category, comprising over half a million reviews. Through manual exploration, we have identified a wide

array of linguistic expressions such as slangs, different abbreviations of words, and spelling errors. One of the example summaries we got was just "Cough Medicine" which shows a small sample of the challenge we were facing with the data. Another issue we ran into was that overall, there were a lot more positive reviews compared to negative reviews, despite the dataset reflecting a vast spectrum of user sentiments and vocabulary (encompasses approximately a quarter million users and over 75 thousand different food products).

To enhance the quality of our training data, we have undertaken a rigorous data-cleaning process. This involves:

- Removing HTML and web artifacts
- Filtering out stopwords using the NLTK library
- Cleaning text with regular expressions to remove non-alphabetic characters and extraneous spaces
- Applying word count thresholds to ensure data consistency
- Utilizing a custom text cleaning function
- Iteratively processing and retaining data that meets our quality standards

This cleaning and standardization of data are crucial to ensure our model is trained on relevant and accurate information. We will provide a detailed visual representation of the dataset's characteristics, such as sentiment distribution and word frequency, in the final report.

4 Model Architecture

We have selected the T5-base model as the foundation for our transformer architecture, inspired by the paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"[3]. The T5-base model is particularly suited for tasks involving variable-length input and output sequences, a common characteristic of review texts. T5's unified text-to-text approach transforms every task into a text generation problem, aligning well with our objective to convert raw reviews into concise summaries. More on the reason for the selection can be found in the discussion section.

4.1 Overview of T5-base

The T5-base model comprises an encoder and a decoder, each with 12 transformer layers. It utilizes a vocabulary size of 32,000 tokens, specifically designed for the T5 tokenizer. Unlike traditional models that use GloVe or BERT embeddings, T5 employs its own pre-trained embeddings, making it adept at handling a wide range of language processing tasks.

4.2 Encoder

The encoder in T5-base processes input sequences using a stack of 12 transformer blocks. Each block contains a self-attention layer followed by a feedforward neural network. The self-attention mechanism allows the model to weigh the importance of different words in the input sequence, enabling it to capture the context effectively.

Each transformer block in the encoder includes layer normalization and a residual connection, enhancing training stability and preventing the vanishing gradient problem. The output from the final encoder block is then passed to the decoder.

4.3 Decoder

The decoder, mirroring the encoder's structure, also consists of 12 transformer blocks. It takes the encoder output and the previously generated output tokens as inputs to generate the next token in the sequence. The decoder employs masked self-attention, allowing each token to only attend to previously generated tokens, which is crucial for autoregressive generation.

Like the encoder, each block in the decoder includes self-attention layers, layer normalization, and a residual connection. The final output of the decoder is passed through a linear layer and a softmax function to generate a probability distribution over the 32,000 token vocabulary, from which the next token in the summary is selected.

4.4 Adaptations and Fine-Tuning

For our specific task of summarizing customer reviews, we will fine-tune the T5-base model on a dataset of review-summary pairs. This fine-tuning will adapt the model's weights to better suit the summarization of reviews, optimizing it for concise and contextually accurate output.

4.5 Advantages of Using T5-base

The choice of T5-base is motivated by its flexibility and effectiveness in various natural language processing tasks. Its unified approach to handling different types of text-to-text transformations makes it a robust choice for generating accurate and coherent summaries of customer reviews.

4.6 Model Architecture Figure

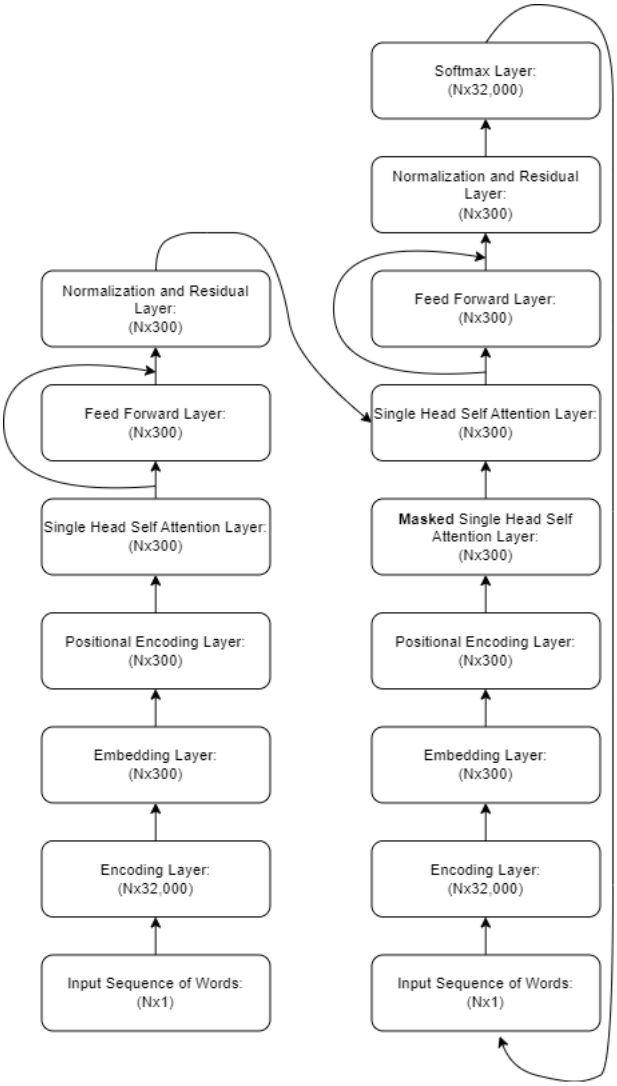

Figure 1: Architecture figure is based on the transformer model introduced in "Attention is All You Need"[1] with the modifications mentioned above.

5 Results

Metric	Trained Model	Baseline Model
BLEU	0.228966	0.18471
ROUGE-1	0.416104	0.24515
ROUGE-2	0.251672	0.011014
ROUGE-L	0.268903	0.017273

Figure 2: BLEU, ROUGE-1, ROUGE-2, and ROUGE-L scores compared to the baseline model.

5.1 Baseline Model

We chose to compare our trained model with the original general purpose T5 sequence-to-sequence model in order to evaluate the effectiveness of our training. The T5 model is known as an industry standard pre-trained model and is a typical choice for comparisons of pre vs post trained sequence-to-sequence transformer models. Furthermore, the T5 model serves as the base of our training model. Therefore, justifiably comparing against this model will verify whether we have made improvements in text summarization, as evidenced by improvements in BLEU, ROUGE-1, ROUGE-2, and ROUGE-L scores.

5.2 Choice of Performance Metrics

The choice of performance metrics mentioned above were justified as they are very commonly used scores to track how well a language model is doing in comparing predicted text to target text. The choice of these metrics were also used in place of accuracy because they paint a more holistic picture of the models performance. More specifically BLEU(Bilingual Evaluation Understudy) was chosen because it measures the number of words in the models output that matches the target while taking into account the order of the words, giving a good understanding of the precision of the model. ROUGE-1 was selected as it captures the extent of which individual words appear in both the generated and target text(as more of a surface level metric). ROUGE-2 was further selected as it calculates bi-gram overlap and is sensitive to the order as well. ROUGE-L was finally chosen as it measures the longest common sub-sequence between the two texts, it measure overall semantic similarity and is not as sensitive to specific words like the other 3 metrics. Overall in combination these four metrics give a holistic picture into how well our model is doing, and was chosen based on it's known reliability as metrics in evaluating language models.

6 Discussion

The metrics presented in 2 offer an in-depth insight into the performance of our trained model compared to the baseline. Analyzing these scores helps us assess the model's success

in generating accurate summaries of customer reviews.

6.1 Comparative Analysis of BLEU Scores

Firstly, the BLEU score achieved by the trained model was 0.228966, a notable improvement over the baseline model's score of 0.18471. This suggests that the trained model's predicted summaries align more closely with the target summaries, particularly in terms of shared words and the precision of word order in the output sequence.

6.2 Evaluation of ROUGE Scores

As shown in 2, the ROUGE scores, which analyze the overlap of n-grams between the target and predicted summaries, further support this conclusion. The trained model achieved a ROUGE-1 score of 0.41604, nearly double the baseline model's score of 0.24515. This significant increase indicates that the trained model captures more keywords from the target summary, regardless of their order.

6.3 Analysis of ROUGE-2 and ROUGE-L Scores

Furthermore, the ROUGE-2 and ROUGE-L scores, assessing longer subsequences and the longest common subsequence between the target and prediction, respectively, also favor our trained model. The ROUGE-2 score of 0.251672 and ROUGE-L score of 0.268903, compared to the baseline's scores, demonstrate its enhanced ability to maintain coherence at the phrase level and preserve sentence structure. This indicates an improved understanding of longer contexts and cohesive structures.

6.4 Overall Assessment of Model Performance

In summary, the trained model demonstrated improvements across all metrics, indicating its effectiveness in generating accurate summaries that capture the essence of the reviews. Compared to a standard industry baseline, our implementation shows marked progress in producing precise summaries.

6.5 Rationale for Choosing T5

Initially, we explored various transformer-based models, including BERT and BART, alongside manual implementations. The strengths and limitations of each model were considered, leading to our final choice.

6.5.1 Limitations of BERT and BART

BERT's strength in contextual understanding was evident, but its limitations in text generation, often resulting in incoherent sentences, made it unsuitable for our summarization needs. BART, while better at generating fluid summaries, fell short in handling extensive reviews effectively.

6.5.2 Advantages of T5

Our choice ultimately settled on the T5 model. T5's text-to-text framework, treating every task as a text generation problem, excelled in summarizing Amazon Fine Food reviews. It adeptly condensed lengthy reviews into concise summaries, preserving key sentiments and information.

6.5.3 T5's Scalability and Adaptability

Furthermore, T5's scalability and flexibility enabled fine-tuning for the specific language and nuances in Amazon Fine Food reviews. Its superior performance in initial tests, combined with its adaptability and high efficacy in various text generation tasks, solidified T5 as our preferred model for review summarization.

6.6 Limitations

Despite the promising performance of our model, we acknowledge several limitations that warrant attention. One notable concern is the quality of certain review summaries generated by the model. In some cases, the summaries lacked coherence or failed to capture nuanced sentiment, possibly due to the inherent complexity and variability of natural language in user reviews. Additionally, the summaries occasionally missed idiomatic expressions or humor, which are challenging for even advanced models like T5 to interpret accurately.

Furthermore, our model's training was limited by the computational resources at our disposal, which constrained the extent of hyperparameter tuning and model experimentation we could feasibly conduct. This limitation might have impeded our model's ability to learn more complex patterns within the data, potentially affecting summary quality.

Another limitation arises from the dataset itself. Despite our extensive cleaning process, some noise in the data—such as non-standard use of language, typos, or grammatically incorrect sentences—remained. These elements could lead to inaccuracies during training and affect the model's ability to generalize from the training data to real-world applications. Some of the summaries themselves in the data set were just the product name which isn't a summary.

Finally, while we have taken steps to mitigate bias by removing any identifiable information and ensuring privacy, there is still the potential for unintended bias in the data that was not fully addressed. Such biases could lead the model to generate summaries that inadvertently perpetuate stereotypes or overlook the diversity of customer opinions.

7 Ethical Considerations

The ethical landscape of automating review summarization is multifaceted, encompassing the entire spectrum from data collection to model deployment. Ethical data handling must ensure privacy and informed consent, and the dataset should reflect diverse demographics to avoid bias amplification. We will conduct a rigorous audit for biases during model training to ensure balanced representation. Concerns extend to post-deployment, where misuse could lead to manipulation of consumer opinions or economic harm.

Moreover, we recognize the potential for our model to inadvertently homogenize consumer feedback, which could diminish the cultural richness and diversity of individual narratives. To address this, we will develop our model to honour the uniqueness of each review while ensuring accurate summarization.

Finally, we will adhere to the ethical guidelines as outlined by the NeurIPS conference, recognizing that ethical machine learning is an ongoing process requiring continuous reflection and adaptation. Our team commits to maintaining an open dialogue about the ethical dimensions of our work and to seeking input from diverse stakeholders to inform our approach.

References

[1] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Łukasz Kaiser Ashish Vaswani, Noam Shazeer and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[4] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:1073–1083, 2017.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.