Summarization and Sentiment Analysis of Reviews

Aayush Indrapratap Singh
North Carolina State University
asingh48@ncsu.edu

Bhavya Omprakash Agrawal
North Carolina State University
bagrawa@ncsu.edu

Diksha Paliwal
North Carolina State University
dpaliwa@ncsu.edu

1 Background and Introduction

Visits to online businesses have surged rapidly as internet access has expanded to remote regions. As a general tendency, people seek suggestions and feedback from the prior customers to determine whether or not to acquire goods online. This project will contain a method for providing users with concise and accurate product reviews.

The project's goal is to create a model that accurately summarizes reviews, which will be useful for both customers and sellers who can use the data to enhance their services and products. It is also taking into consideration two kinds of people, one sample who gives a low rating with a negative review and others who give a low rating for an overall positive review. We intend to overcome this disparity with our summarizing and sentiment extraction.

Furthermore, reading long reviews requires more time and effort, but these long reviews contain the maximum information. Most individuals simply overlook these helpful reviews due to their length. Because of this, the buyer might not know about all the aspects including the positive as well as the negative points, that they should be considering, before buying the product. People will be able to get the information more quickly if they were provided with the summarized reviews.

Doing this would allow the customer to read less data but still gain the most important information needed to decide on the product. It would also improve productivity by speeding up the surfing process of the user.

So far, there are a lot of research papers related to product reviews, sentiment analysis or opinion mining. For example, Xu Yun et al. [2] from Stanford University applied existing supervised learning algorithms such as perceptron algorithm, naive bayes and supporting vector machine to predict a review's rating on Yelp's rating dataset. They used hold out cross validation using 70% data as the training data and 30% data as the testing data. The author used different classifiers to determine the precision and recall values. Callen Rain [1] proposed extending the current work in the field of natural language processing. Naive Bayesian and decision list classifiers were used to classify a given review as positive or negative. Deep-learning neural networks are also popular in the area of sentiment analysis.

2 Method

2.1 Encoder-Decoder LSTM seq2seq model

The Encoder-Decoder architecture is mainly used to solve the sequence-to-sequence (Seq2Seq) problems where the input and output sequences are of different lengths. The objective is to build a text summarizer where the input is a long sequence of words (in a text body), and the output is a short summary (which is a sequence as well). So, we can model this as a Many-to-Many Seq2Seq problem.

There are two major components of a Seq2Seq model:

1. Encoder
2. Decoder

Long Short Term Memory (LSTM) are capable of capturing long term dependencies by overcoming the problem of vanishing gradient.

2.1.1 Encoder

An Encoder Long Short Term Memory model (LSTM) reads the entire input sequence wherein, at each time step, one word is fed into the encoder. It then processes the information at every time step and captures the contextual information present in the input sequence. The below diagram which illustrates this process:

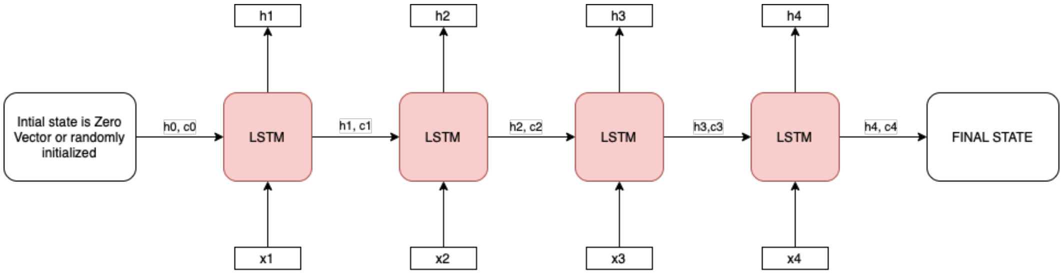


Figure 1: Encoder Model

The hidden state (hi) and cell state (ci) of the last time step are used to initialize the decoder. Remember, this is because the encoder and decoder are two different sets of the LSTM architecture.

2.1.2 Decoder

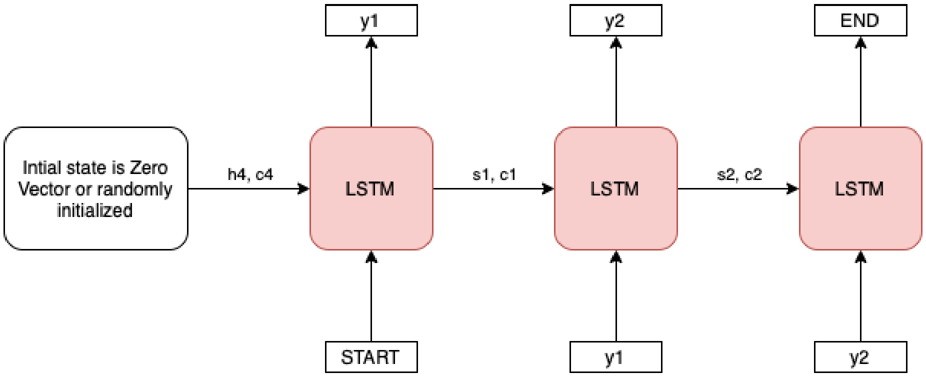


Figure 2: Decoder Model

The decoder is also an LSTM network which reads the entire target sequence word-by- word and predicts the same sequence offset by one time step. The decoder is trained to predict the next word in the sequence given the previous word.

<start> and <end> are the special tokens which are added to the target sequence before feeding it into the decoder. The target sequence is unknown while decoding the test sequence. So, we start predicting the target sequence by passing the first word into the decoder which would be always the <start> token. And the <end> token signals the end of the sentence.

2.2 Sentiment Analysis

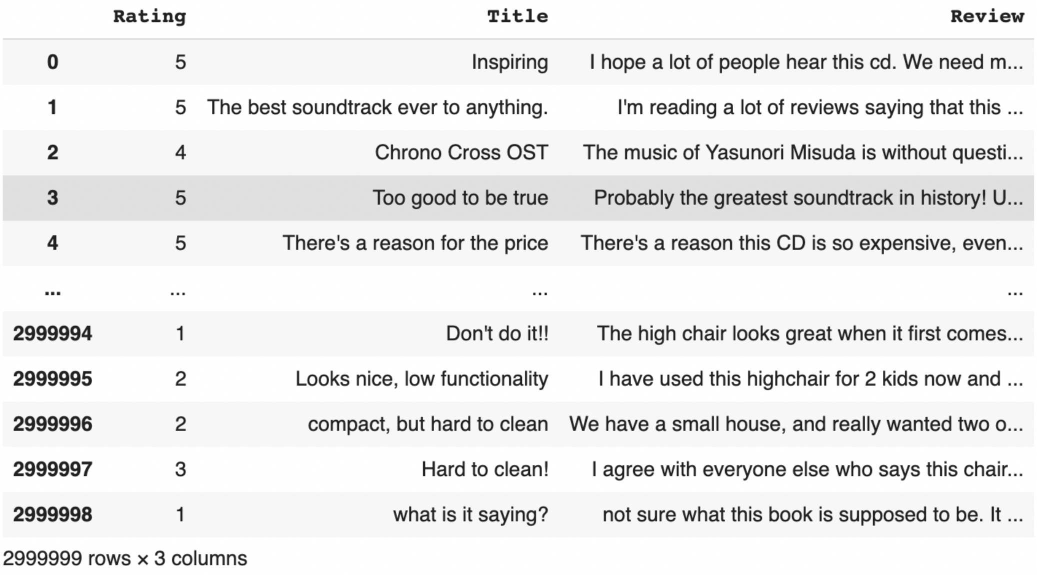
Sentiment analysis models detect polarity within a text (e.g. a positive or negative opinion), whether it's a whole document, paragraph, sentence, or clause.

Understanding people's emotions is essential for businesses since customers are able to express their thoughts and feelings more openly than ever before. By automatically analyzing customer feedback, from survey responses to social media conversations, brands are able to listen attentively to their customers, and tailor products and services to meet their needs. The service and the product review's polarity is the rating the user provides for that review. The Good Reviews are those with rating 5 stars, 4 stars and 3 stars and Bad Reviews are those with rating 2 stars and 1 star. Finally, when a feature sentiment is extracted the sentiment phrase is sent to a polarizer method, this method basically returns 1 if the phrase is a positive sentiment else 0 if the phrase is a negative sentiment.

3 Experimental Setup

3.1 Dataset

The Project uses an Amazon reviews dataset from Kaggle (link to dataset) which contains the rating, title and reviews of approximately 3 million amazon customer reviews in the training set and approximately 600,000 amazon customer reviews in the testing set.



	Rating	Title	Review
0	5	Inspiring	I hope a lot of people hear this cd. We need m...
1	5	The best soundtrack ever to anything.	I'm reading a lot of reviews saying that this ...
2	4	Chrono Cross OST	The music of Yasunori Misuda is without questi...
3	5	Too good to be true	Probably the greatest soundtrack in history! U...
4	5	There's a reason for the price	There's a reason this CD is so expensive, even...
...	...	...	...
2999994	1	Don't do it!!	The high chair looks great when it first comes...
2999995	2	Looks nice, low functionality	I have used this highchair for 2 kids now and ...
2999996	2	compact, but hard to clean	We have a small house, and really wanted two o...
2999997	3	Hard to clean!	I agree with everyone else who says this chair...
2999998	1	what is it saying?	not sure what this book is supposed to be. It ...

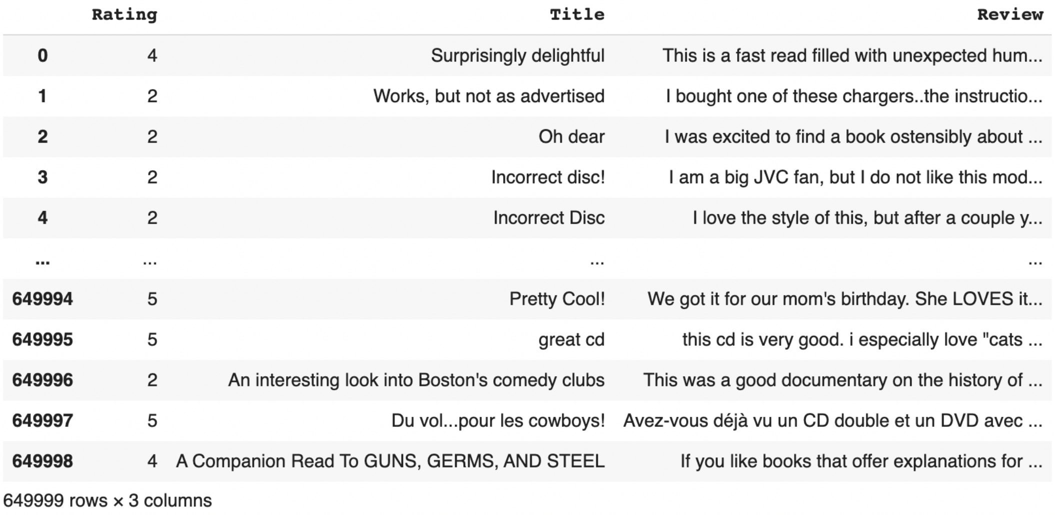
2999999 rows × 3 columns

Figure 3: Training Dataset

3.2 Preprocessing

The project is utilizing natural language processing techniques to preprocess the data into the format needed for sentiment analysis and summarization.

1. **Convert all text to lowercase**: The lower() method returns the lowercase string from the given string. It converts all uppercase characters to lowercase.
2. **Removing all punctuation**: Format words and removing unwanted characters.
3. **Tokenization**: Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. Long reviews are converted into lists of words to further summarize and find the polarity.
4. **Removing stop words**: Removing unnecessary words by using English stop words imported from spacy-de library. Stopwords include any word in the sentence which do not provide any meaningful insight to the project's data analysis.
5. **Lemmatization**: Procuring base word or lemma from the inflected forms. Lemmatization is preferred over stemming since lemmatization provides the root word without changing the



	Rating	Title	Review
0	4	Surprisingly delightful	This is a fast read filled with unexpected hum...
1	2	Works, but not as advertised	I bought one of these chargers..the instructio...
2	2	Oh dear	I was excited to find a book ostensibly about ...
3	2	Incorrect disc!	I am a big JVC fan, but I do not like this mod...
4	2	Incorrect Disc	I love the style of this, but after a couple y...
...	...	...	...
649994	5	Pretty Cool!	We got it for our mom's birthday. She LOVES it...
649995	5	great cd	this cd is very good. i especially love "cats ...
649996	2	An interesting look into Boston's comedy clubs	This was a good documentary on the history of ...
649997	5	Du vol...pour les cowboys!	Avez-vous déjà vu un CD double et un DVD avec ...
649998	4	A Companion Read To GUNS, GERMS, AND STEEL	If you like books that offer explanations for ...

649999 rows × 3 columns

Figure 4: Testing Dataset

meaning of the word whereas in stemming the root word might not have the same meaning as the original word.

4 Results

1. Preprocessing and cleaning of the dataset has been done.
2. The models are under construction.

5 Conclusions

We've finished pre-processing our dataset and turned it into a format that can be used for summarization and sentiment analysis models. We're currently investigating and learning about different summarization methods that can help us get better results. We'll run the sentiment analysis model on the summarized reviews after the summarizing model is finished. This will identify the polarity of the reviews, allowing us to compare and contrast how the ratings and polarity of the reviews differ.

References

[1] Callen Rain. 2013. Sentiment analysis in amazon reviews using probabilistic machine learning. *Swarthmore College* (2013).

[2] Yun Xu, Wu; Xinhui, and Wang; Qinxia. 2015. Sentiment Analysis of Yelp‘s Ratings Based on Text Reviews. (2015).

A Appendix

Original Plan: Dataset to be taken using Web Scrapping

Current Plan: As per the advised received from the TA, dataset for the project is taken from Kaggle.