Purdue University
Computer Science
CS57700: Natural Language Processing

Less is More: Human and AI Summarization for Sentiment Analysis

Spring 2022

Pablo Tomás Campos	campos72@purdue.edu
Laura Martínez	mart2180@purdue.edu

Table of Contents

1 Introduction

Sentiment Analysis is of great importance for e-commerce. With companies having to analyze great bodies of text, it would be cheaper to analyze summaries of said texts.

However not all bodies of text come with a summary provided by the user. Could this summary be efficiently produced by a computer rather than by a human?

This is the topic of our project: compare the results of a sentiment analysis after being trained with the body of reviews and after only being trained by their generated reviews by our abstractive summarization model. As a baseline, we can train a third model on the text summarization target, the human written summaries of the reviews. We believe this could improve the results of the sentiment analysis task, by reducing the size of our vocabulary to only its most meaningful words.

2 Research Question

The object of this project is to evaluate if sentiment analysis tasks can be improved by using summaries of the texts in question. Moreover, we will also asses if this the difference in performance between human and AI generated summaries in order to see if the process of summarization can be automated for a sentiment analysis task.

3 Modeling

This project is composed of two main sections: building an abstractive summarization model from reviews and building sentiment analysis classifiers from the original texts, its human written summary and the summary generated from the abstractive summarization model previously mentioned.

As for the classification task, our literary review [1] has lead us to conclude the following: a L2-regularized logistic regression on bag of words representations serves as a good baseline in order to compare classification models. We can also conclude that both LSTM and biLSTM can be considered state of the art architectures for sentiment analysis when trained on top of word embeddings such as Glove or word2vec, or even randomly initialized embeddings.

Following up with the abstractive text summarization, we have used the following two papers by Google: the first one presenting the transformer architecture [4] and the second one presenting the BERT model [2], which excels in text summarization. In search for a lighter transformer, we finally ended up using the T5 model [3].

Even though we have found extensive literature for both comparing sentiment analysis models and abstractive text summarization, we have been unable to find any previous work done resembling our intentions for this project. This incites us even greater to carry out this project, as we will have an opportunity to venture and pioneer into the unknown.

4 Dataset

For this project we need a dataset with two golden labels: first, the headline or summary for a given text and sentiment score or ranking. Luckily enough, the Amazon Fine Food Reviews dataset matches perfectly to our needs.

This dataset contains reviews primarily of fine foods, but it also includes reviews from all other Amazon categories. In total it contains 568,454 reviews dating from October 1999 to October 2012. In order to produce results faster, we will only be working with a stratified sample of 50,000 reviews in order to preserve the distribution of attribute *Score*

From this dataset, we are only interested in the following three attributes:

1. **Text**: The review left by the user on a certain product. Our most important piece of data, since it will be used to train both the abstractive text summarization model and the sentiment classification model.
2. **Summary**: A brief summary of the review, made by the same user that wrote the review. It will be used as the target when training the abstractive summarization model, as well as to train a sentiment classification baseline model.
3. **Score**: Rating of the product on scale of 1 to 5 left by the same user who wrote the review. Will the used as the golden label of our sentiment classification task.
4. **Generated Summary**: Not present in the original dataset. Corresponds to our AI generated summary from *Text*.
5. **Label**. Not present in the original dataset. Remapping of attribute *Score* as explained below.

The distribution of attribute *Score* is the following:

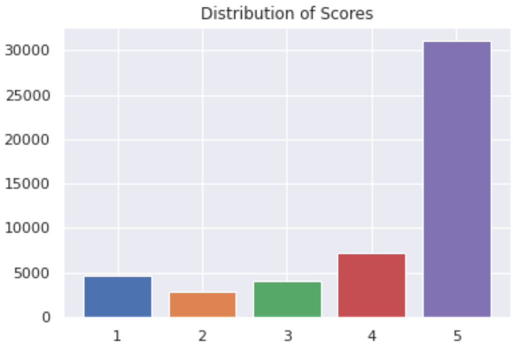


Figure 1: Distribution of attribute *Score*

Because scores 1 and 2 do not have many instances compared to score 5, we have decided to map the 5 values to 3 into a new attribute *Label*: scores 1 and 2 will become ”negative”, score 3 will become ”neutral” and scores 4 and 5 will become ”positive”. This union of scores helps the sentiment analysis models, as it increases the available training data per label

We will train the two sentiment analysis models previously exposed: a bow logistic regression and a LSTM with randomly initialized word embeddings, to predict the attribute *Label* separately on the following attributes of the dataset: *Text*, *Summary* and *Generated Summary*

5 Experimental Evaluation

In this section we shall propose both the model used for the abstractive text summarization and for sentiment analysis based on the related work cited on said section.

As for the abstractive text summarization, we will be fine-tuning a pre-trained transformer, the state of the art architecture for NLP tasks.

For the sentiment classification task two models will be compared: a L2-regularized logistic regression trained on a bag of words representations of the training examples to serve as our baseline; and a LSTM trained with randomly initialized embeddings of the training examples.

Moreover, we will also both models on the summaries of the reviews generated by fine-tuning a pre-trained a transformer model. The quality of the summaries generated by this transformer will be judged using the rouge score on a summary level. As for the sentiment classification models, since the values of the attribute *Score* are not uniformly distributed across its range of values, precision, recall and the f1-score will be used to asses their performance.

6 Experimental Results

6.1 Abstractive Text Summarization

For evaluating the results of fine tuning a pretrained transformer trained on attribute *Text* to predict attribute *Summary* of the dataset we used the following variations of the rouge metric:

- Rouge-N: Measures the number of matching n-grams between the target and the generated sentence. We used both rouge-1 and rouge-2.
- Rouge-L: Measure the longest common sub-sequence between the target and the generated sentences. The idea behind this metric is that longer shared sub-sequences are indicative of high similarity between to sentences.

The obtained metrics and loss across the 10 epochs of training are presented below:

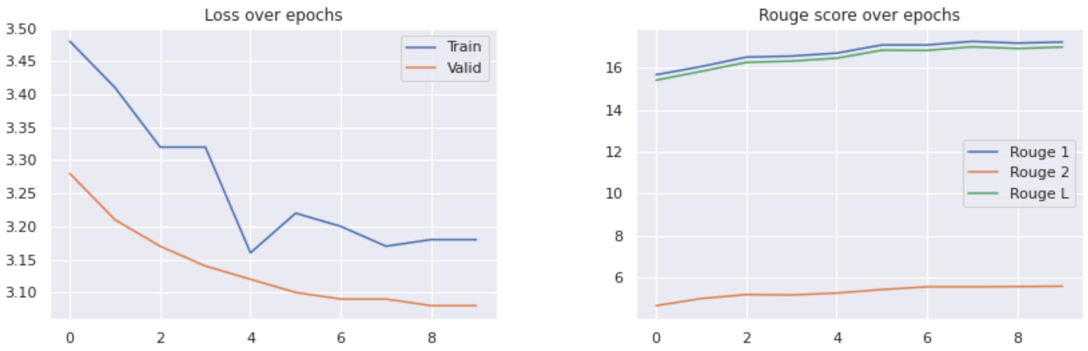


Figure 2: Loss and metrics over training epochs

And here we can illustrate some the human written summaries, corresponding to attribute *summary* together with the models generated summary, corresponding to attribute *generated summary*.

Summary	Generated Summary
Good Quality Dog Food	Great product
Not as Advertised	Jumbo Salted Peanuts - I had this!
"Delight" says it all	The most flavorful treat
Cough Medicine	Best product in the industry
Great taffy	great taffy

6.2 Sentiment Analysis

In this section we shall analyze the performance of both our models architectures, bow logistic regression and randomly trained embedding LSTM, on the three different attributes: *Text*, *Summary* and *Generated Summary*. All the metrics and confusion matrices have been computed by a test set composed of 10% of the available data, that is 5000 entries.

6.2.1 Logistic Regression

Here we present the metrics obtained by the three logistic regression models trained on the attributes *Text*, *Summary* and *Generated Summary* respectively:

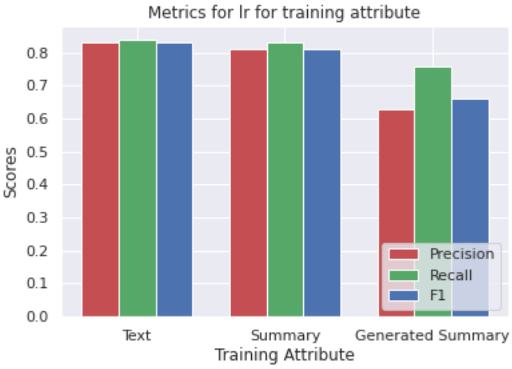


Figure 3: Metrics for the three logistic regression models

Contrary to our initial hypothesis, the logistic regression model trained on attribute *Summary* did not perform better than the one trained on attribute *Text*, despite using a bow feature representation that favors smaller vocabularies as *Summary*. Moreover, the logistic regression model trained on *Generated Summary* performed significantly worse than its other two competitors.

See apendix A. It is also to be noted that the logistic regression model trained on attribute *generated summary* did not predict a single time the neutral sentiment and it barely predicted the negative sentiment. This implies that either the bow feature representation of the generated summaries or the generated summaries themselves do not capture the underlying sentiment of the review.

6.2.2 LSTM

As for the lstm, hyper-parameter tuning was conducted using a validation set of 10% of the remaining 45000 training instances. The hyper-parameter tuning was conducted by measuring the validation loss, and ceasing training when it had not improved its best result for 10 epochs, after which the model was to be returned to its best state for validation loss.

Here we present the metrics obtained by the three logistic regression models trained on the attributes *Text*, *Summary* and *Generated Summary* respectively:

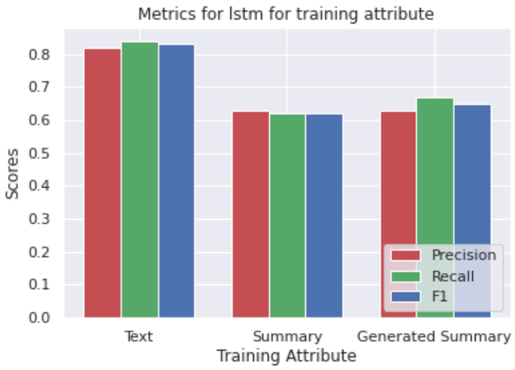


Figure 4: Metrics for the three lstm models

The results obtained from the three lstm models trained with randomly initialized embeddings of attributes *text*, *summary* and *generated summary* are more aligned with our initial hypothesis that the results offered by the previously mentioned logistic regression models. We can observer how, despite the lstm trained on attribute *text* is still outperforming its other two competing models, the lstm trained on *generated summary* outperforms the remaining lstm trained on attribute *summary*.

From these results we can imply that there is improvement by substituting human written summaries for AI generated summaries, as the lstm performs better on the latest. However, both human and AI summaries fail to perform better than the lstm trained on the complete review. Out initial hypothesis thus results to be only half true: sentiment analysis models seem to prefer the whole text rather than its review, however they also prefer AI generated reviews rather than human generated reviews.

A Confusion Matrices

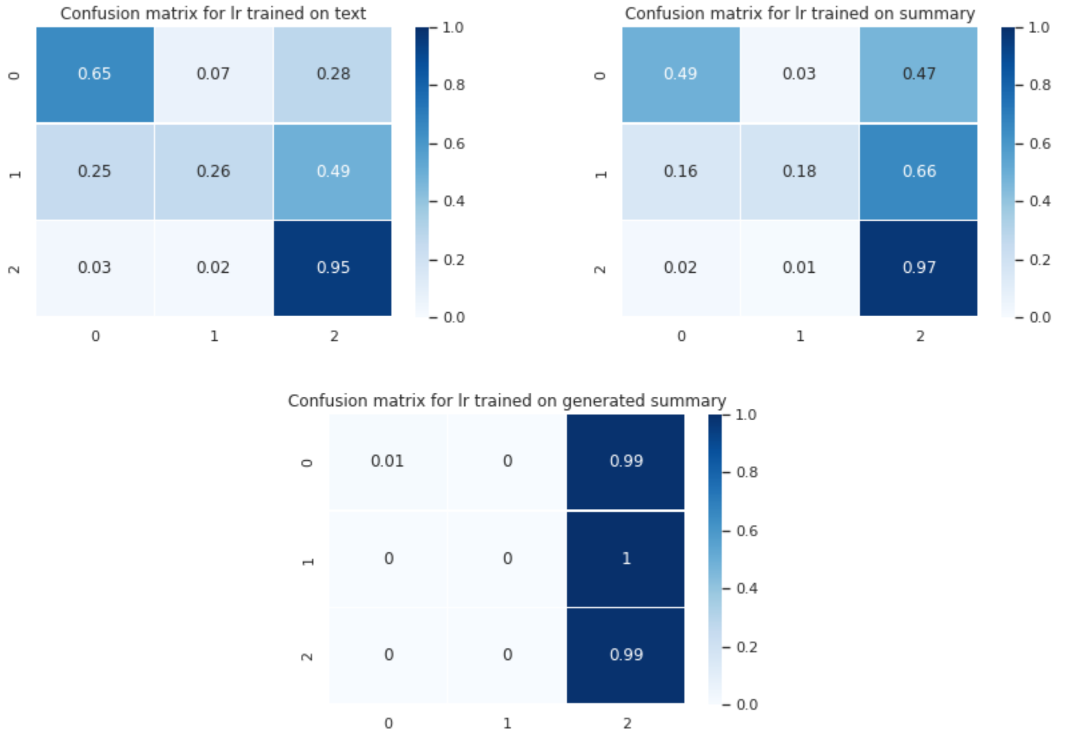


Figure 5: Confusion matrices for the three logistic regression models trained

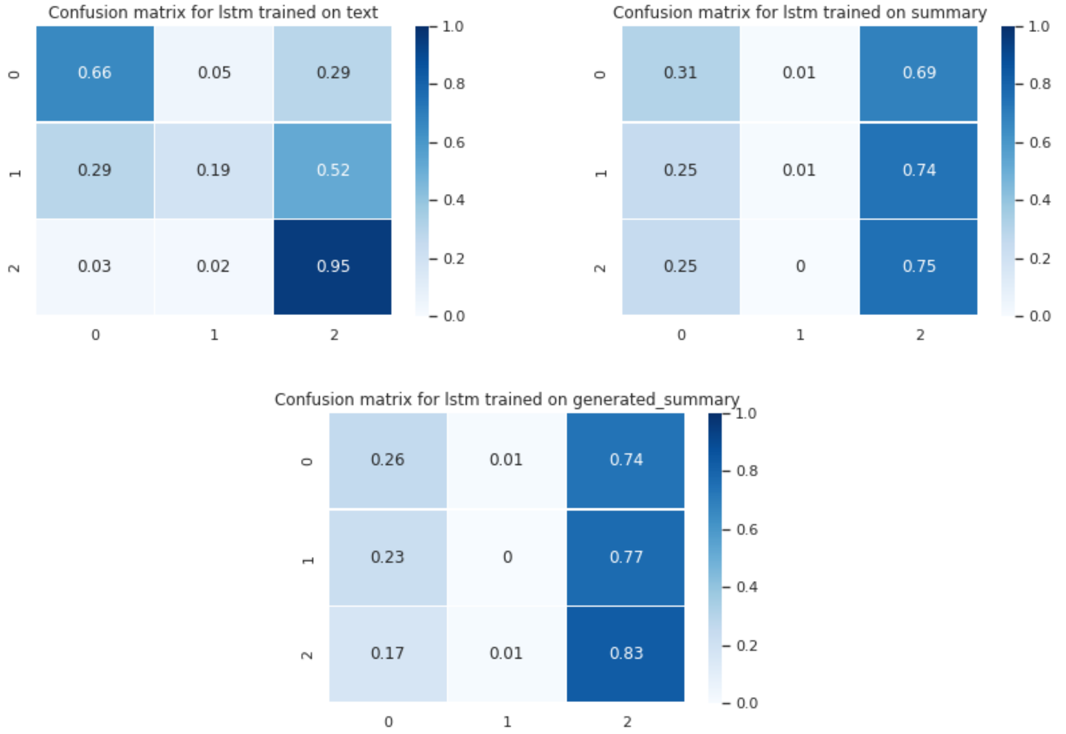
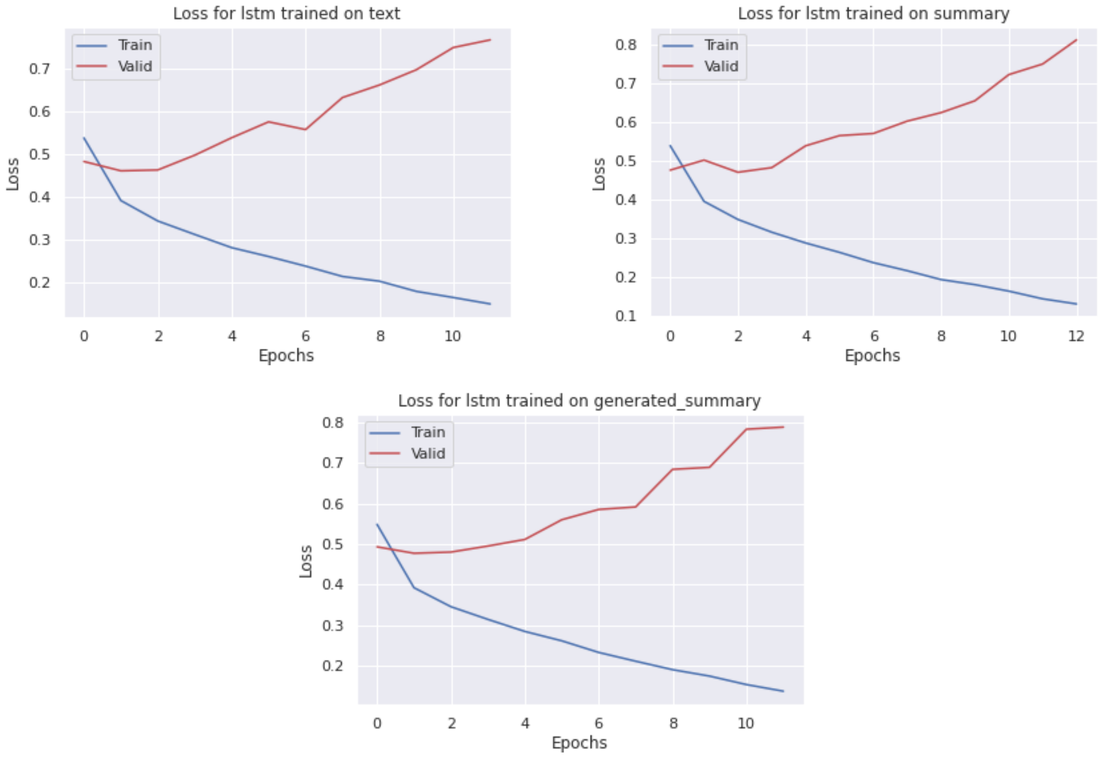


Figure 6: Confusion matrices for the three lstm models trained

B Loss Curves

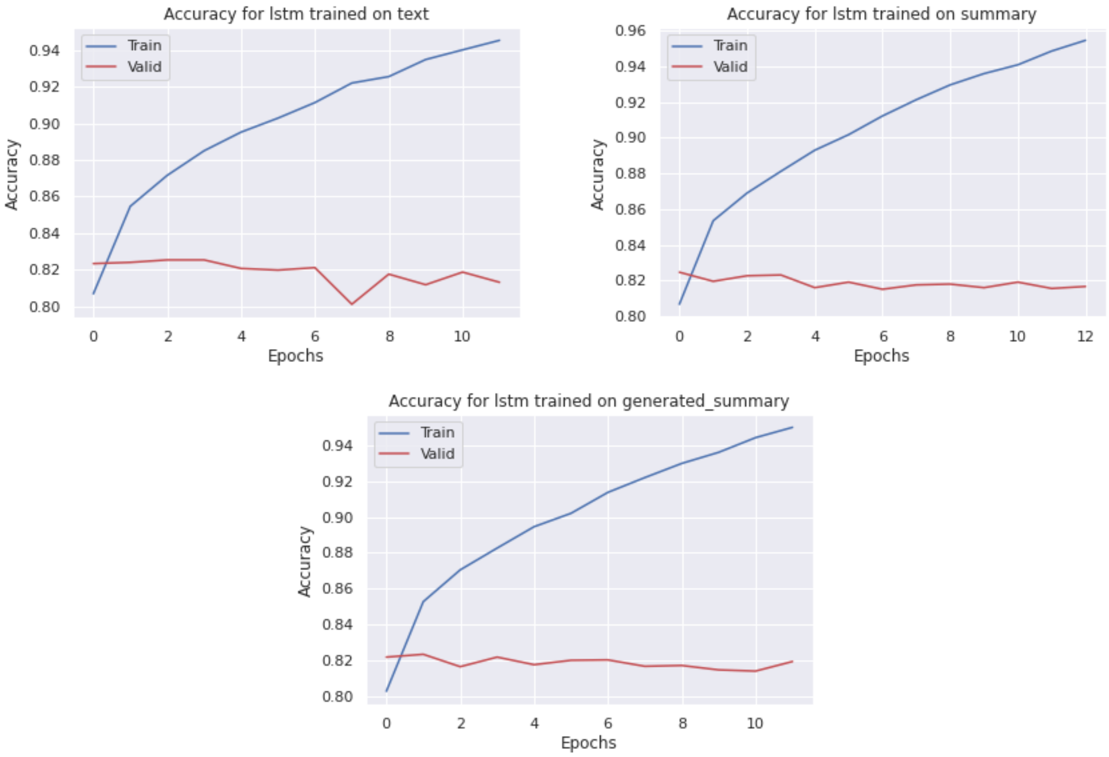

Loss for lstm trained on text

Loss for lstm trained on summary

Loss for lstm trained on generated_summary

Figure 7: Loss curves for the three lstm models trained

C Accuracy Curves


Accuracy for lstm trained on text

Accuracy for lstm trained on summary

Accuracy for lstm trained on generated_summary

Figure 8: Accuracy curves for the three lstm models trained

References

[1] Jeremy Barnes, Roman Klinger, and Sabine Schulte im Walde. Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 2–12, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2–12, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[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. *CoRR*, abs/1910.10683, 2019.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.