Purdue University
Computer Science
CS57700: Natural Language Processing

Less is More: Human and AI Summarization for Sentiment Analysis

March 2022

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

Table of Contents

1 Problem Formulation and Motivation

With the boom of text generative models with the excellent results obtained by the GPT3 language model, it seems more than likely that many tasks will start adopting this technique.

One said task could very well be text summarization, where for an input text provided, a summary or headline is generated. Historically this tasks was primarily solved using extractive text summarization, in which the most relevant terms of a sentence are extracted to from a summary. On the contrary, abstractive summarization generates a summary from scratch, even using words that where not even present on the original text.

With covid-19 and the consequent rise of online shopping, there is a big market potential for tools that are able to improve the ends customer experience. Again, text summarization could greatly aid e-commerce by summarizing customers reviews to a sentence, reducing the amount of data to be processed.

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 Background

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.

Following up with the abstractive text summarization, we have used the following two papers by Google: the first one presenting the transformer architecture [3] and the second one presenting the BERT model [2], which excels in text summarization.

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 unkown.

3 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.

4 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 BERT model designed by Google, which uses the state of the art architecture for NLP tasks: transformers.

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 word2vec embeddings of the training examples.

5 Experimental evaluation

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, made by a total of 256,059 users about 74,258 products.

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.

We will train the two sentiment analysis models previously exposed: a bow logistic regression and word2vec LSTM, to predict the attribute *Score* on the following attributes of the dataset:

1. **Text**: The review in question. Will be used as a baseline to see if there is an improvement with using its summary:
2. **Summary**: Summary of the review in question.

Moreover, we will also both models on the summaries of the reviews generated by fine-tuning a pre-trained a BERT 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 Work Plan

This project will be carried out by Pablo Tomás Campos Fernández and Laura Martínez. Both members will be involved in all stages of the project. However, it is to be noted that Pablo Tomás Campos Fernández will take the lead in finetuning the pretrained BERT model, while Laura Martínez will focus her efforts on training the sentiment classification models. The analysis and comparison of the resulting sentiment classification models shall be done by both members of the team.

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