The Positive, the Negative, and Everything In Between:

Enhanced Sentiment Prediction in Movie Reviews Using Deep Learning with the SST-5 Dataset and ChatGPT

Final Project Documentation,

Course: Deep Learning at AIT Budapest,

Authors: Ethan McFarlin and Wolff Gilligan,

Semester: Spring 2023

— Table of Contents:

— Introduction:

Compounded by the profusion of tweets, product reviews, op-eds, and other snippets of subjective information scattered around the web, public opinion takes center stage in the modern information landscape. Coming in all shapes and forms, people's subjective attitudes, value judgements, and sensibilities actively shape our online experience— from the world of personalized advertising to the automated ranking of movie recommendations. But bearing in mind the overwhelming amount of online information characterized by some underlying sentiment, an important question arises: how do we effectively parse through a sample of text and make sentiment-based inferences from the grammar, word choice, sentence structure, or other factors. Enter the world of sentiment analysis, a broad field with applications in natural language processing, machine learning, and other disciplines [1].

This paper will approach the field of sentiment analysis from the more specialized lens of deep learning, engendering a discussion of topics such as dataset selection, model architecture, and the enhancement of AI-based predictions with Chat GPT.

— Preview Solutions:

Existing literature in the AI field has corroborated the viability of models such as deep neural networks for sentiment analysis on datasets including online Twitter feeds [2]. However, analysis in this space is not confined to a select subset of text-based samples, and instead, extends itself across a diverse range of input types such as aggregated movie reviews in the form of datasets like SST-1, SST-2 and MR [3]. In spite of their wide breadth and focus on novel machine learning methods like SVM, NB and ME, existing sentimental analysis papers are lacking in their connection to new language models such as Chat GPT. Moreover, the use of Chat GPT to enhance existing sentiment prediction represents a novel avenue for contribution to the field— a niche which this paper explores in greater length.

— Dataset:

As previously mentioned, a model that can reliably predict the sentiment of a text sample has a myriad of use cases, and this has led to the development of datasets specifically designed for this task. However, the goal of creating a general sentiment prediction model is broad, and often corporations and researchers have

a very narrow use case. This encourages them to create datasets that yield from a specific source. Another incentive here is that since the definition and boundaries of the sentiment of a text sample in one context shifts when within a different context, by keeping the source consistent this helps to clarify the patterns in the data, improving the accuracy of the model. As a result there exist a wide variety of groupings of sentiment prediction datasets tailored for predicting the sentiment of certain types of text, such as movie reviews, product reviews , Twitter posts, car and hotel reviews, and others.

We then shifted our focus over to which dataset we wanted to train our model on, and for a couple of reasons, we decided to go for the SST5— Stanford Sentiment Treebank 5— dataset. The SST5 dataset is sourced from movie reviews, and is one of the largest datasets available, with around 10,000 elements. Since we are mostly interested in ChatGPT's labeling, not in predicting the sentiment of a specific context of text, we reasoned that a larger dataset is more valuable. Second, this dataset's sentiment label is on the sentence level, not on the phrase level, which is critical if we want to further label our data using ChatGPT. Phrase level sentiment is much harder to integrate with ChatGPT, since the sentiment calculations are done on the whole sentence, which would be very hard to communicate to ChatGPT. Additionally, unlike the general SST dataset, SST5's labeling is not binary, but is on a 5-degree sentiment scale. We hypothesized that ChatGPT's labeling would have more of an impact on the accuracy of the model if there were more degrees of sentiment, as it could maybe reflect more patterns that the model initially missed in the second stage of training.

— Proposed Method:

Regarding the choice of model, we gravitated towards a BERT (Bidirectional Encoder Representations for Transformers)-based method, due to its versatility in classifying the meaning of text and easy integration with our tech stack in Python and Keras Tensorflow [4].

Equipped with a preferred selection of dataset and proposed model, we proceeded onto the pre-processing stage. We began by iterating over the SST5 dataset and subsetting it into separate data structures for training and validation (for both the sentences and their underlying sentiments). Our second objective was to perform stop-word removal, which consists of stripping away non-essential

words— those which do not uniquely contribute to the classification of sentiment— from the input dataset. We curated a custom selection of stop-words which we felt bore special relevance to the realm of sentiment analysis in movie reviews.

At this stage, we imported a pre-trained BERT model, defined a max sentence length based on the dispersion of the data, and performed tokenization across the input data. This step ensured that each word was associated with a unique numerical identifier. Further prepping the data for model fitting, we constructed a Tensorflow dataset from our encodings and shuffled the training/test subsets.

With the necessary pre-conditions in place, we compiled and fitted our BERT-model with an Adam optimizer using sparse categorical cross entropy as the loss function. We measured its performance across a predetermined number of epochs and reported the accuracy.

Having fit the model to the SST5 dataset, our focus now pivoted over to the second phase of analysis: enhancement of sentiment predictions using ChatGPT. Our broad theoretical approach was as follows: we would use the OpenAI API to query ChatGPT for sentence completions. For each input sentence, we appended an instruction to the input, instructing the model to return a sentiment analysis and familiaring it with our ranking system. In this manner, we generated new sentiments for each of our previous sentences.

The same methodological approach carried out on the original dataset (i.e. stop-word removal, BERT-based tokenization, etc.) was applied to the new, enhanced predictions. Finally, we further trained the existing model on the additional data with the goal of improving its accuracy through the introduction of a wider variance of sentence labels.

— Results and Discussion:

Since we first train the BERT-based model on the data where the labels are sourced from the dataset itself, and then train it on the ChatGPT-labeled data, we have two stages of evaluating the accuracy.

The first stage had a final epoch accuracy of 47.6%. While this is worse than random for a binary classification question, our sentiment is rated on a 5-point scale, and therefore is a regression problem. This means that our model actually performed much better than random (which would have an accuracy of 20%, since it has a ⅕ chance of guessing correctly).

This result is consistent with the literature, as the choice of dataset, preprocessing steps, and model are all proven steps in the problem of sentiment prediction, and so the results are unsurprising. The dataset is extensive and tailored specifically for this problem, as aforementioned, and is one of the most popular choices for this type of prediction. The preprocessing is pretty standard for text processing, as it makes sense to tokenize and remove the stop words, whereas it would not make sense to shuffle the order of the words within each sentence. The BERT-based model, while somewhat new (being developed by Google and released in 2018), is widely used for creating models based on text, as it creates a context for a word in a sentence. Since sentiment is a facet of language that is almost completely defined by context, this model is an amazing choice.

Unfortunately, GPT 3.5 introduced Rate Limits on API requests, and at 60 a minute, we cannot relabel our entire dataset using ChatGPT. However, we were able to relabel 30 sentences with ChatGPT. We were able to keep the labels consistent with our base model, as in on the same 5-point scale, by giving it a custom prompt within the query.

However after retraining, it did slightly decrease the accuracy of the model/ This makes sense since the new dataset was too small for the model to adjust accurately, so it was effectively like adding noise.

Overall, our BERT-based model was able to somewhat accurately predict the scale of the sentiment of sentences sourced from the SST5 dataset.

— Works Cited:

1. Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat and A. Rehman, “Sentiment Analysis Using Deep Learning Techniques: A Review” International Journal of Advanced Computer Science and Applications(IJACSA), 8(6), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080657
2. A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," 2017 7th International Annual Engineering Seminar (InAES), Yogyakarta, Indonesia, 2017, pp. 1-4, doi: 10.1109/INAES.2017.8068556.

3. A. Aslam, U. Qamar, P. Saqib, R. Ayesha and A. Qadeer, "A Novel Framework For Sentiment Analysis Using Deep Learning," 2020 22nd International Conference on Advanced Communication Technology (ICACT), Phoenix Park, Korea (South), 2020, pp. 525-529, doi: 10.23919/ICACT48636.2020.9061247.
4. Rani Horev. “BERT Explained: State of the Art Language Model for NLP.” Medium, Towards Data Science, 10 Nov. 2018, towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.