Scoring and Summarization of Movies/Products Reviews

Ashish Peruri
University of Illinois at Chicago
vperur2@uic.edu

Sai Teja Karnati
University of Illinois at Chicago
skarna3@uic.edu

Srikanth Maganti
University of Illinois at Chicago
smagan20@uic.edu

ABSTRACT

Sentiment Analysis and Summarizing of reviews, a popular application problem, has gained a lot of attention this decade. Most of these products sold on e-commerce sites have thousands of reviews given by customers. Customers interested in a product have a right to know the product better without going through thousands of reviews. We want to provide a sentiment, Unique rating(Machine Critique Score) and a neat summary (Machine-generated review) of the total reviews. We use transfer learning over BERT and some other models like LSTM, GRU, Bidirectional LSTM to generate the sentiment of the products. A different modified unique rating score is also created. We work on the well established Centroid based Extractive Summarization method to get an overall review of products. The results and models are studied to reflect on learnings.

Keywords

Sentiment Analysis, Bert, LSTM, GRU , Transfer Learning, Extractive Summarization

1. INTRODUCTION

Rapid changes and great research efforts in the field of NLP can be seen recently. Deep learning has drastically transformed the practice of NLP over the last decade. The recent research on Transformers and BERT are turning heads with significant results across multiple NLP tasks. BERT leverages concepts from some existing approaches including ELMo and ULMFiT. The core advance with BERT is that it masks different words in any given input phrase and then estimates the likelihood of various words that might be able to fill that slot. The resulting contextual pre-trained embeddings were outperforming many state of the art models at that time besides reducing the cost and complexity of training substantially.

Sentiment analysis or opinion mining has been an area of interest for almost all major businesses. It helps the companies to understand their customer's opinions and improve their experience. Potential customers also would like to the common opinion and a neat summary of the product/service before purchasing it. The information gained from product reviews can be useful for researchers in better market prediction.

However, saying this, the task of providing a good opinion and summary remains formidable because of the proliferation of data. Many major sites typically contain huge volumes of data containing lots of biased, redundant and opinionated text. The average human being will have difficulty identifying relevant information and summarizing the information in it. Any challenge that is difficult for human beings is a complex task for computers as well.

In this paper, we will be providing sentiment of each product in a huge dataset, either positive or negative based on reviews from thousands of customers. We are also providing a Unique rating which uses the ratings and helpfulness feature available in the dataset. Along with the sentiment and unique rating, we are providing a summary of the product as understood from all the product reviews using the popular centroid based extractive summarization method while exploring the abstractive summarization approaches [3]. The dataset used in this paper are Customer reviews of around 10000 products and a total combined reviews of 200000. We compare the accuracies obtained from transfer learning using BERT pre-trained embedding with LSTMs, Bi-LSTMS, and GRUs in the sentiment analysis task. We will evaluate the goodness of generated summary in the summarization task with the ROUGE score using CNN-DailyMail dataset/opinosis dataset with golden standards.

2. RELATED WORK

Transfer learning from pre-trained embeddings has a rising interest in the field of NLP. We use the BERT model for transfer learning in Sentiment analysis task along with some other LSTM models. We use the large BERT model during Summarization task. We briefly review the related approaches in the Pre-training, Sentiment Analysis, and Summarization.

2.1 Sentiment Analysis

Sentiment Analysis understands the contextual and hidden meaning of the given text. There has been incredible research being done in this area with around 7000 papers published already. Sentiment analysis was mainly done on public opinion during the 20th century. Since the beginning

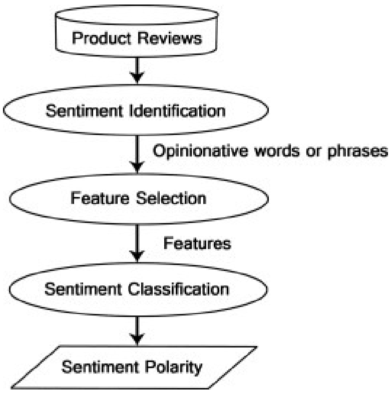

Figure 1: Sentiment Analysis Process

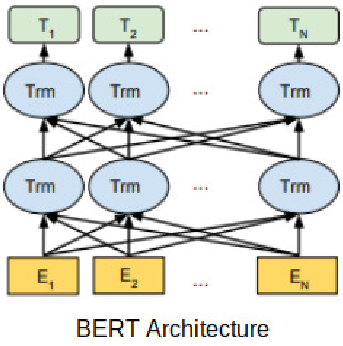

Figure 2: BERT Architecture, where E_n is the n-th token in the input sequence,Trm is the transformer block, and T_n is the corresponding output embedding.

Table 1: $BERT_{BASE}$ vs $BERT_{LARGE}$

Hyperparameter's	$BERT_{BASE}$	$BERT_{LARGE}$
No of Layer's	12	24
No of hidden units	768	1024
No of self-attention heads	12	16
Total trainable parameter's	110M	340M

of the 2000s, text analysis which was only being done by Computational Linguistics Community from the beginning of the 1990's had a big outbreak due to an abundance of text data from the web.

Deep learning has gained a lot of attention recently as the model started outperforming many Machine Learning algorithms such as Naive-Bayes and KNN. Some significant work has already been done on sentiment analysis using Vanilla RNN, LSTM and GRU using glove word embeddings.

We will be using the BERT model for pre-trained embeddings and compare the performance against other models like LSTM,BiLSTM [1] and GRU.

2.2 Pre-trained Language Models

Pre-trained word embeddings are a crucial part of the recent NLP models as they often give significantly better results (Mikolov et al., 2013; Pennington et al., 2014) than those embeddings learned from scratch on a rather small domain-specific dataset. Generalizations of Word embeddings like sentence embeddings (Kiros et al., 2015; Logeswaran and Lee, 2018) and paragraph embeddings (Le and Mikolov, 2014) are also used in some models.

The recent methods in pre-training language models trained on a large network with a large amount of unlabeled data and fine-tuned over specific tasks made a breakthrough in many NLP tasks, such as OpenAI GPT and BERT (Devlin et al., 2018). BERT is pre-trained on the Masked Language Model task and Next Sentence Prediction task via a large cross-domain corpus. Unlike biLM limited to a combination of two unidirectional language models, BERT used a Masked language Model to predict words that are randomly masked or replaced. This paper uses BERT fine-tuning method and demonstrates it's great potential in transfer learning.

2.3 Summarization

Initial work on the summarization task started almost at the same time the field of Natural language processing started. In 1958, some papers on the statistical method of scoring and selecting sentences from larger blocks were published. One of the popular models demonstrated compelling results. At the same time, both metrics such as ROUGE and datasets as DUC were developed for comparing and contrasting the various methods for summarization. However, most of this work has been in the form of extractive text summarization.

Abstractive text generation has been a recent area of interest for many researchers in this field. Recent work, from various research groups at Facebook and IBM, have built models that combined the task of extracting with that of generating by using recurrent neural networks. RNNs usually with the generation of the seq2seq model can both extract valid information and generate valid language using the same computational framework. They have shown State-of-the-art performance on Sentence Summarization task using a feed-forward window-based neural network with an attention mechanism. In our case, we attempt to build, contrast and understand simple extractive and abstractive summaries [3].

3. PROBLEM STATEMENT

With the increasing number of user-generated reviews, it has become an excruciating task to process the most relevant feedback from the reviews without machine interference. In this work, we aim at providing each Product/Movie with : a Sentiment(either positive or negative), MCS (Machine Critique Score), and review summary: MGR (Machine Generated Review) using different models while exploring different options.

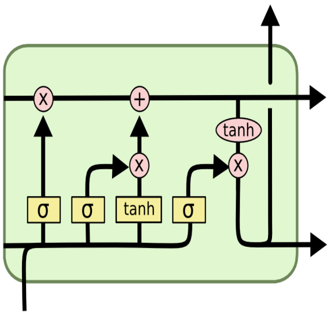

Figure 3: LSTM Architecture

4. TECHNICAL APPROACH

Our approach to this project can be divided into two subsections - understanding of baseline models and implementing our advancements to the existing model. In this paper, we explore different models to contrast against our base model using Bert. Importance and impact of pre-trained embeddings form bert are studied as well.

4.1 LSTM Networks

Long Short Term Memory is based on RNN model built to avoid exploding and vanishing gradient. LSTM consists of cell, input gate,output gate and forget gate in which the cell remembers the specific value in particular time period and the remaining gates take care of flow of information and manages this by learning when to remember and when to forget.LSTM is vastly used in the areas of text classification.

4.2 Bidirectional LSTM Networks

In case of Bidirectional LSTMS, We feed the Algorithm with the data which actually iterates from beginning to end and end to beginning with simulataneous tagging to sequence. We have access to both past and input features for a given time.So, Bidirectional LSTM network works as shown in figure 7.It helps us in making use of future features(via forward states) and also past features(via backward states) in a certain time period.

4.3 BERT

We use BERT(Bidirectional Encoder Representations from Transformers) base case model for contextual pre-trained work embeddings and use them as part of transfer learning to a smaller domain specific dataset using fully connected feed forward layer. We even use the

4.4 BERT LSTM

With the objective of extending an already impressive model into more successful model. We added a LSTM over the BERT base model instead of simple fully connected layer.

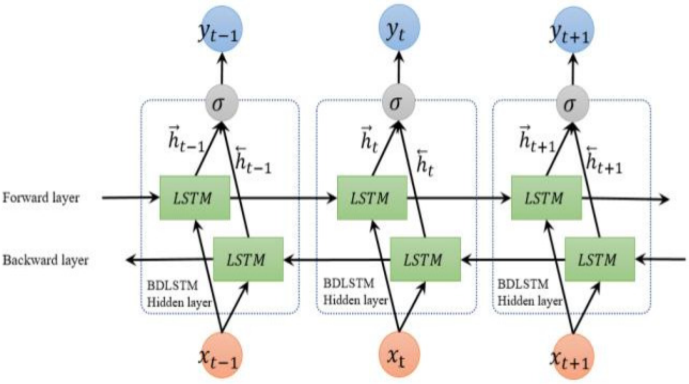

Figure 4: Bidirectional LSTM Networks Architecture

Table 2: Statistics about the Data

statistics	overall(Gaming)	overall(Beauty)	overall(Toys)
count	231780	198502	165797
mean	4.09	4.19	4.36
std	1.20	1.17	0.99
min	1.0	1.0	1.0
25%	4.0	4.0	4.0
50%	5.0	5.0	5.0
75%	5.0	5.0	5.0
max	5.0	5.0	5.0

4.5 Extractive Summarization

For the summarization task, a well known method,Centroid based Extractive [2], is used. The Review/Text is transformed into tensors/vectors which will be used to calculate centroids using K Means algorithm. Based on the Compression/Summarization ratio, we get the closest sentences/phrases that convey the overall message of the reviews. We used large uncased Bert model to create the tensors for our extractor.

4.6 Machine Critique Score

Machine Critique Score(MCS) is calculated using the ”helpful” feature in the Amazon Product Reviews dataset which conveys the importance factor of that particular review . The normalized function with alpha(def $= 0.5$) gives a rating of x additional frequency based on the helpfulness*alpha. This helps reviews which convery truth as they have more helpful points compared false reviews.

5. EXPERIMENTAL SETUP

In this section, we expand on the data used, preprocessing, methodology, experimental details (e.g., hyperparameters) and evaluation metrics.

5.1 Data

Our Data set contains consumer reviews of Amazon products(taken from Consumer Reviews of Amazon Products

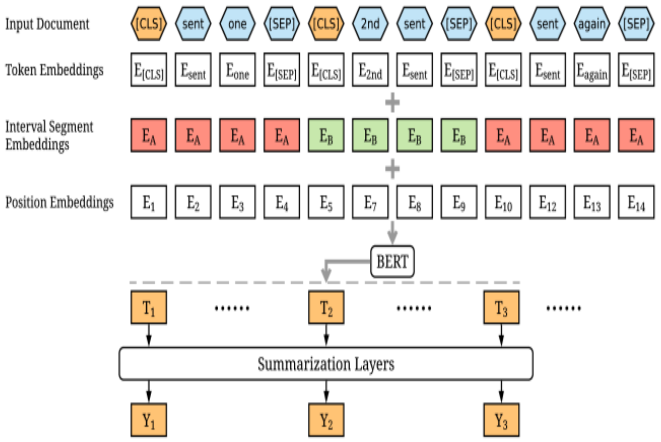
Input Document
[CLS] sent one [SEP] [CLS] 2nd sent [SEP] [CLS] sent again [SEP]
Token Embeddings
E[CLS] Esent Eone E[SEP] E[CLS] E2nd Esent E[SEP] E[CLS] Esent Eagain E[SEP]
Interval Segment Embeddings
EA EA EA EA EB EB EB EB EA EA EA EA
Position Embeddings
E1 E2 E3 E4 E5 E7 E8 E9 E10 E12 E13 E14
BERT
T1 T2 T3
Summarization Layers
Y1 Y2 Y3

Figure 5: Extractive Summarization Architecture

Data set). This repository basically has around 143 million reviews which were collected spanning May 1996- July 2014. Three categories that we worked on are Video games, Toys, and Beauty care. Each example has similar features such as type, name of product, review, rating of the product and some related metadata like descriptions,category,etc.

5.1.1 *Data Pre-processing*

We perform the following Pre-processing steps on the review text before we feed into our model.

1.First, we converted everything into lower-case letters,then removed all the unnecessary jargon like digits,punctuation symbols through by regular-expression form.

2.Removing all the metadata that won't be used like product/user information.

3.Creating a new column that contains concatenated summary(provided by Amazon in review summary column) of all reviews for each product. This will be used during the summarization task.

5.2 Experiments

We explore each task individually. We expand on transfer learning from BERT, LSTMs, GRUs, and BiLSTM models. We will analyse the hyper-parameters used and the training data from Amazon review dataset. Different tasks require different features and different parameters.

5.2.1 *Sentiment Analysis task*

In the sentiment analysis task, we use the amazon review dataset with features like review text. We don't have a class information of the samples to train, so we use the rating to create class based on rating, positive if average rating >= 3.5 and negative otherwise. We use this to train our model so that it could predict the sentiment on the test data.

We initially tried transfer learning using a base-Uncased-BERT model. We added 2 feed forward layer of size 768 nodes and parameters as part of the fine tuning. We obtained decent result of 87.63. Then we append a LSTM layer to the BERT pre-trained model. We observe a slight increase in the accuracy. Further improvising the network added to the pre-trained model may help in increasing the accuracy.

In order to compare and contrast the results, we also used a LSTM model, GRU model and Bidirectional LSTM model. Surprisingly, these models gave decent results.We used a fixed sequence length of 150 equivalent to the average length reviews.

5.2.2 *Machine critique score Task*

Machine critique score is our simple exploration towards new ways to score products rather than simply averaging the score. We use the helpfulness column along with ratings column to calculate this new score. We used a small alpha value to smooth the scores since most of the helpfulness column had values=0 as most reviews don't get upvoted for being helpful.

5.2.3 *Summarization Task*

In the Summarization task, we use the Amazon dataset with features like review text and review text summary. Instead of using all long reviews of each product, we created a column that contains concatenated summary of all reviews which will be used to create a product summary.

We work with both 'bert-base-uncased' and 'bert-large-uncased' models. The text to be summarized initially is fed to the model where it is tokenized using 'Bert-tokenizer'. The bert model uses contextual pre-trained embeddings, which give us some tensors that contains the information of the text to be summarized. We use these send this information into the Extractor class where it uses the Kmeans clustering algorithm to finds the centroids. Based on our compression ratio, we get the closest tensors from each centroid and convey the information in a compressed fashion. Our approach is popularly known as Centroid based Extractive Summarization.

We tried both 'bert-large-uncased' and 'bert-base-uncased' and the 'bert-large-uncased' gave slightly more sensible summaries. The model had a hidden size of 1024, 24 hidden layers and 16 attention heads. We used a drop out of 0.1 for attention and hidden layers. It takes around 1 minute to generate 30-40 summaries. The text to be summarized had around 500-1000 words which was compressed to around 50-100 words since we used a compression ratio of 0.1.

This is an example of concatenated review of Super Smash Bros. : *"an amazing.....good multiplayer game....great fighting game for n. fun strategic and fast.... old school".* It

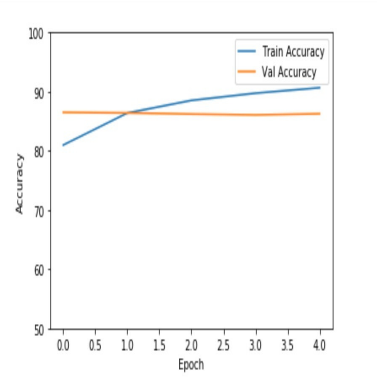

Figure 6: Accuracy vs epochs

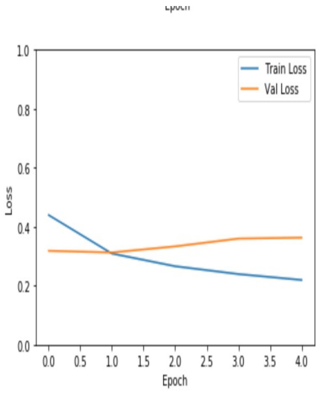

Figure 7: Loss vs epochs

had around 377 words, 2000 characters. The summary for this was given as *"lets smash something then be really bored. one of the best and must have games for the n console. if you own a nintendo then you should own this game. smash em other fighter games this game rox. my kids play this game more than any other."* The summary contains 40 words and contains the essential information that it is good, fun and old school fighter game. This is extracted summary containing sentences/phrases from the main text and some sentences/phrases might be a little off in terms of context flow. We will be working on abstractive summarization in the future to create summaries that can be generated instead of being extracted.

6. RESULTS

We evaluated the sentiments of products based on the review over the Video games, Beauty products and Toys/games datasets as shown in Table 3. We achieved around 87-88% in the Bert base case, 85% in both BiLSTM and LSTM, 88-89% in the case of LSTM over Bert model.The transfer learning over base Bert model improved the performance by 2-3% which can be further improved by fine tuning the Bert pre-trained embdeddings using the domain specific dataset. Instead of Bert Base case, Bert Large case can be used given enough resources for better results. LSTMs seem to perform decently. BiLSTM gained 1% in slight performance over LSTM. GRUs seem to perform way worse than others. Summaries from the Centroid based Extractive Summarization method seem to encapsulate overall information about the product. We noticed some redundancy issues in some summary parts. A compression/summarization ratio of 0.1 has successfully reduced the text of size 400-500 words to 40-50 words which is around 3-4 lines of summary. The Machine Critique score, MCS, gave more meaningful rating than the normal rating as the normal averaged rating system considers highly helpful rating and false/unhelpful rating as similar. We can improvise upon this unsupervised summarization by making it semi-supervised. This can be done by first gene

Table 3: Accuracy for different models.

Model	Video Games	Beauty	Toys and Games
LSTM [4]	85.4	83.8	85.7
BiLSTM	85.5.4	84.1	85.91
GRU	83.3	80.1	83.6
BERT	87.63	85.2	88.14
BERT-LSTM	88.23.6	87.4	89.6

7. CONCLUSIONS AND FUTURE WORK

We have utilized transfer learning on BERT for Extractive Summarization and Sentiment Classification. For Sentence Classification, we tried LSTM, BiLSTM, GRU, BERT, and BERT-LSTM. From the results, we observe that BERT-LSTM performs best on the test data. In the future, we can improve directly from the reviews instead of just the helpful variable, and we can also include spam filtering in the summarization pipeline to achieve a better summary. With the increasing online reviews, we believe such methods would become instrumental in deciding the review of a product. We are currently working on adding a Attention layer to the pre-tained BERT Model to improve upon sentiment classification. Also we want to extend our current unsupervised Extractive summarization to semi supervised Abstarctive summarization.

8. TEAM WORK DIVISION AND OVERALL EXPERIENCE

All authors contributed equally in the project. The course Deep learning in NLP with this project work inspired us to look into various areas of research related to NLP. A lot of papers and articles were read while exploring different models and possibilities.

9. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[2] Dragomir R. Radev, Hongyan Jing, Magorzata Sty, and Daniel Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40:919–938, 2004.

[3] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, 2015.

[4] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*, 2012.