Point-of-Interest Review Topic Modeling via Semi-supervised Learning

Yucong John Hu
yh2860@nyu.edu

Heeseok Joo
hsj278@nyu.edu

Zexi Ye
zy1311@nyu.edu

Abstract

In this paper, we will outline how our team, in association with mentors from TripAdvisor (NASDAQ: TRIP), tested several machine learning models on unlabeled, raw online user reviews on points of interest (POI) to produce a satisfactory accuracy result on 4-class topical classification. For the final output, we not only predicted a single label class, but also the distribution of probability over the 4 classes to provide more granular data and to allow levels of thresholding.

Our best-performing model is the Bidirectional Encoder Representations from Transformers (BERT), while other out-of-the-box supervised learning machine learning models achieved lower, but still commendable result. A completely unsupervised approach using both latent Dirichlet allocation (LDA) and Guided LDA failed to produce any clusters of interest to us, even with seed words enforcement.

We envision various business use cases and future improvements with our result and BERT model. One of the constraints, however, is the lack of ground truth in the user review data set. This means that we were unable to produce an independent assessment of test accuracy because even in cases where we manually labeled the review set, there were disagreements amongst teammates with regard to which class does a review sentence belong to, and the labeled data is inherently subjective and subject to the bias of 3 people who labeled it.

1 Introduction

1.1 Background

Since 2010s, research in the field of natural language processing has been extraordinarily active and a wide range of firms have been relying on state-of-the-art techniques to solve practical business problems, benefitting significantly from the progress. Briefly, natural language processing (NLP for short) describes the class of problems associated with raw text/audio through modeling. Examples of downstream tasks include sentiment analysis, language modeling, machine translation, etc.

Traditionally, NLP problems were largely tackled by rule-based programs or statistical models such as the autoregressive model in language modeling. Over the past decade, however, deep learning has been the trend in regard to NLP tasks. Notably, neural network-based models often prove to be superior given a sufficient amount of training data and a carefully tuned model configuration. Further, general-purpose NLP models such as the Bidirectional Encoder Representations from Transformers (BERT), featured in Google's 2018 paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, have shown extraordinary performance regardless of specific downstream tasks.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.2 Our use case

Working with our mentors from TripAdvisor (NASDAQ: TRIP), we created a topic-based multiclass classification model using the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT), achieving a validation accuracy as high as 0.7748. The business motivation behind this project is to classify user reviews into 4 classes: emotional (E), factual (F), tips (T), and others (O) on a sentence basis.

As the website that serves travelers worldwide on their big and small adventures, TripAdvisor would want to provide the most relevant, up-to-date information on points of interest to website visitors in an automated fashion. Our result is significant because it is much higher than the out-of-the-box supervised machine learning models that we will outline later. Also, the model is confident on cases where it produces correct probability distribution on the label classes, meaning that it makes a good business use case where the company can highlight the most confidently-predicted reviews from users on its website.

Working with our partners at TripAdvisor, we have achieved a satisfactory performance score and we believe what we have produced is applicable to the company's business model. Namely, by enabling the company to highlight relevant information about points of interest based on sustained and updated user inputs aligns well with TripAdvisor's value proposition to its users and customers.

2 Data understanding

2.1 Overview

We were initially given 100,000 rows of user reviews of varying sentence lengths, where 100 points of interest (POI) in total are included and each POI is given 1,000 reviews. One of the most significant obstacles we encountered so far is the fact that all review data are unlabeled. Given the massive volume of data, it is unrealistic to manually label all of them, so we need to approach the problem diplomatically. We later manually labeled 3,000 rows of user review sentences, and received labeled data categorized as either 'factual', 'emotion' or 'tips', which is not user reviews.

For the initial stage of data pre-processing, we parsed the user reviews into sentences as we were only interested in classifying each review sentence. We observed that due to the randomness in which different users compose their reviews and also the parsing/tokenizing library that we use, sentences with length less than 3 contains mostly punctuation. We therefore dropped them as we proceed with data cleansing. In the end, we were left with about 793,997 user review sentences.

For the training and validation data, we manually labeled 3,000 user review sentences as aforementioned and then randomly sampled from the labeled 'factual' or 'tips' data to produce a balanced distribution over the four classes. As we divided the manual labeling process amongst three teammates, we agreed on strict guidelines on labeling to ensure consistency in our labeling. We also randomly doubled checked manually labeled data to make sure everyone agreed with the labeling. In the end, it gave us around 4,100 rows of labeled data (both manually labeled user reviews and labeled non-reviews provided by TripAdvisor) sentences to work with.

2.2 Caveats

While we were manually labeling, we noticed cases of inconsistency where each of the teammates disagree over how to label a given sentence. We were able to impose stricter definitions and standardize the process. Still, the process of manual labeling by three teammates is inherently subjective and maybe prone to bias shared by the persons who are labeling.

The motivation behind over-sampling from the labeled 'factual' or 'tips' data is that in a 'real-world' data set, we would see disproportionately high numbers in the 'emotional' and 'others' class while the 'factual' and 'tips' class are under-represented. We observed such trend when we were manually labeling the 3,000 rows of the dataset, and concluded that albeit being a reasonable phenomenon, the natural imbalance in distribution would degrade the performance of any models due to a lack of learning examples. Therefore, we proceeded with over-sampling from the 'factual' and 'tips' dataset that was provided to outside the original 100,000 user reviews.

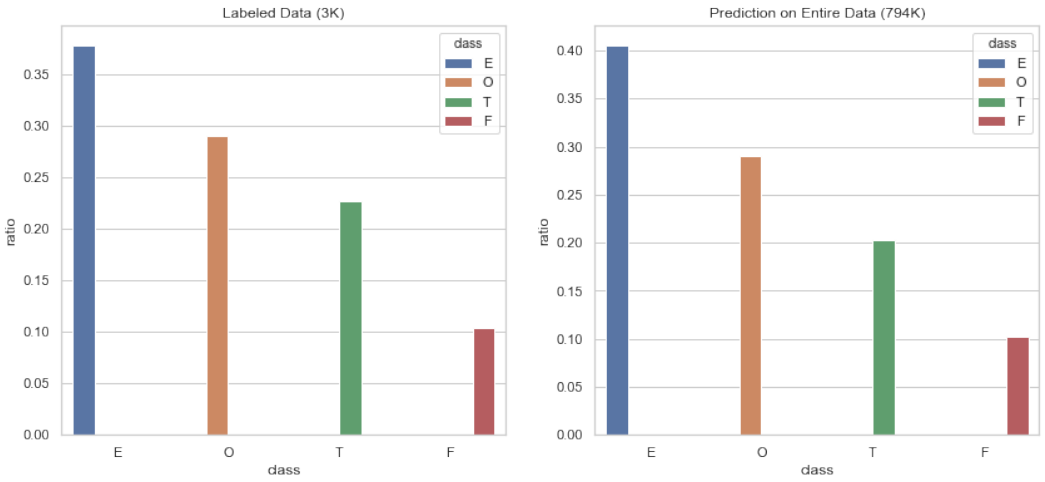

Figure 1: Class distribution plots

3 Methodology

Our approach to this project is one of rapid prototyping to produce a working, baseline model and improve upon baseline in each subsequent iteration. To this end, we divided the responsibility of modeling among 3 team members, each focusing on LDA & guided LDA, naive supervised models, and BERT. We then each tried to produce a working baseline model for each and worked towards incremental improvement.

From the onset, it turned out that our best baseline model as well as the one with the highest potential is BERT. As discussed earlier, deep learning has been gaining traction rapidly over the past decade when it comes to NLP. It is therefore natural to see BERT outperform its peers in supervised and semi-supervised learning.

After deciding on BERT as the best candidate, we then moved forward with hyperparameters tuning and producing outputs that are relevant to the business goal of the company. By trying different models on the same train and test dataset, our methodology shows BERT's advantages over other models.

4 Modeling

For machine learning models, we envisioned using supervised, semi-supervised, and unsupervised models to find the best performance result and business use case for our project.

4.1 Latent Dirichlet Allocation (LDA) & Guided LDA

4.1.1 LDA

In order to separate each reviews into categories based on relevant words, we used Latent Dirichlet Allocation (LDA), which is an unsupervised topic modelling technique that picks out what topics the reviews can be grouped together based on the relevance of the words within the reviews [1]. In LDA, each review is related to words that appear in it, and LDA tries to finds a pattern where certain reviews are connected to the same word thus giving them more relevance to that word and to each other. However, with such large dataset, the connection is often too numerous, so latent layer is used to reduce the number of connections. Latent layer is basically a group of words that is the superset of more specific words, such as time is superset of words like hour or day.

The implementation of classic LDA proved itself to be not useful because the model groups words into categories that it thinks are related regardless of the topics that we selected (Facts, Emotion, Tips) As it can be seen from the figure of top words from each category, it is difficult to justify the grouping since a lot of words are often ambiguous or even unrelated to the assigned topic. Since this

Category	Top 20 Resulting Topic Words
Topic 0	see, visit, go, time, museum, tour, view, one, take, place, great, get, well, day, walk, building, must, beautiful, many, experience
Topic 1	get, go, time, ticket, day, ride, line, park, wait, queue, hour, take, visit, one, would, tour, long, minute, buy, u
Topic 2	get, park, go, ride, take, people, one, good, see, like, food, walk, place, great, time, show, around, also, lot, make

Table 1: Topic words produced by regular LDA

is an unsupervised modelling, there is no specific goal other than the number of categories that we want these words to fall into.

4.1.2 Guided LDA

Since we have set of categories in mind, it was necessary to use Guided LDA, which is semi-supervised learning that guides the words into the categories based on previously given seed words that are related to each topic with chosen level of confidence in these words [2]. In our case, as we were labeling our data, we noticed a pattern between different categories and which words might be appropriate for seed words. For category of 'Emotion', it was by far the most flexible category for choosing seed words since any adjective to describe emotion could be used. For 'Tips', we realized that there are commanding verbs for suggesting other users to take certain actions when they are at the point of interest, so the words that are used when giving advice or suggestion were used. Finally, the category of 'Facts' was the most difficult category to produce since few words can suggest that the entire review is about facts and information about the point of interest. We used words like 'history' or 'year' that could indicate the historical information of the location, or 'tall' or 'meter' that could indicate the physical size or shape of the point of interest.

Category	Seed words used	Top 20 Resulting Topic Words
Facts	history, meter, long, tall, wide, year, old	ride, go, time, park, visit, get, see, people, one, day, like, experience, would, year, take, place, really, make, good, tour
Emotion	interesting, fantastic, breathtaking, fun, beautiful, terrible	see, visit, view, go, great, place, walk, one, take, time, get, museum, beautiful, building, well, tour, must, around, day, also
Tips	aware, recommend, tip, advice, suggest, try, advise, free	get, ticket, go, time, day, line, park, take, ride, wait, hour, queue, one, buy, tour, would, long, minute, people, pay

Table 2: Topic words produced by Guided LDA

Guided LDA returns somewhat more reasonable words compared to what we get from regular LDA, but there is still ambiguity in a lot of words especially for the topic 'Facts'. It makes sense because as we have discussed in the data exploration section, factual reviews are often rare and it is hard to select specific single words that make reviews to be considered factual. Qualitatively speaking, the topic of emotion was the most accurate among the three topics since any type of adjective that describes visitors' emotion can be used, and the topic of tips had some main keywords such as 'wait' or 'queue' that indicates a suggestion or advice to potential visitors.

After training the model on raw data and testing it on labeled test data set, the accuracy turned out to be around 25.5%, which is disappointingly low to a point where it is basically guessing randomly. In the future, we expect higher accuracy if the parts of speech is labeled or known since we have noticed that the category 'tips' often contains a lot of imperative verbs for suggesting new visitors to take certain action for better visit.

Category	Seed words used	Top 20 Resulting Topic Words
Facts	history, meter, long, tall, wide, year, old	visit, see, museum, tour, time, go, one, place, experience, great, would, well, people, like, really, make, history, get, take, many
Emotion	interesting, fantastic, breathtaking, fun, beautiful, terrible	see, view, walk, go, visit, take, get, place, great, one, beautiful, around, time, building, park, city, area, also, good, day
Tips	aware, recommend, tip, advice, suggest, try, advise, free	ticket, get, go, day, time, park, buy, take, line, tour, one, food, pay, u, price, would, queue, people, book, good
Others	go, get, see, like, one, take, building, really, ride, back, well, around, many, place	ride, go, get, time, park, day, wait, hour, line, one, take, queue, visit, people, minute, see, long, would, like, back

Table 3: Topic words of 4 categories produced by Guided LDA

When we were labeling our data, we have mentioned that there were reviews that did not fit into any of the categories. Since Guided LDA still returns topic words that are not completely related to each topic, and the ambiguous and vague words that were produced by Guided LDA taken out from the top topic word list and grouped together into a new seed category called 'Others'. After rearranging the seed words and training it with 4 categories, it ended up giving around 23% accuracy, which is worse than running the original 3 category model. We suspect that the lower score was caused by not seeing the underlying connection made by LDA and the relationship between the words and the reviews. For instance, some words may have seemed ambiguous by itself, but was often used in a sentence that definitely fell into one the 3 categories in mind.

4.2 Conventional Supervised Models

4.2.1 Motivation behind conventional supervised models

In selecting the best supervised models as a benchmark, we run 4 popular machine learning classification models out-of-the-box in a 5-fold cross validation, and produced the mean and distribution of the accuracy score in each iteration. The 4 algorithms of choice are

1. Random forest
2. Linear SVC
3. Multinomial naive bayes
4. Logistic regression

These four models are constructed using the sci-kit learn library implementation, and fitted to the data using their default hyperparameters.

The motivation behind using conventional supervised learning as a baseline is that, in the industry, there always exists the alternative of hiring third-parties to conduct manual labeling of data and train a supervised learning model on the labeled data. While such an approach is the most straightforward and less engineeringly expensive, it is difficult to extrapolate the trained algorithm to future use cases; it also exposes proprietary data and accrus extra costs in retaining services such as Amazon Mechanical Turk. Consequently, we have decided to explore this less technically savvy approach by manually labeling some of the data ourselves, and fit the aforementioned 4 algorithms to the labeled data. We will then compare the results of these conventional models to that of BERT, which uses pre-trained embeddings. This way, we hope to highlight the power of using a deep learning model like BERT that is more transferable and has better performance.

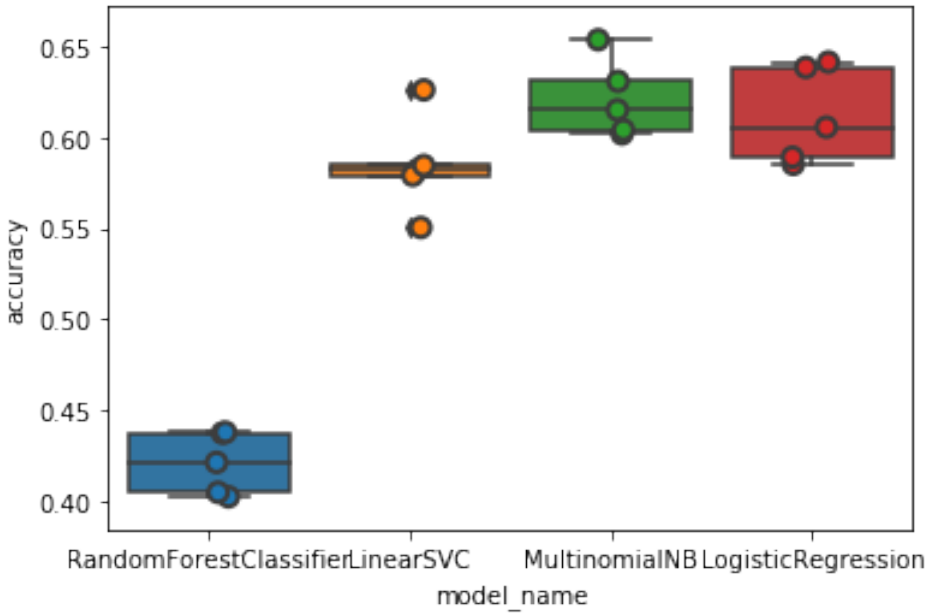

Figure 2: 5-fold cross validation performance across 4 models.

4.2.2 Performance of conventional supervised models

Of the 4 models showcased above, we can see that multinomial naive bayes has the best mean performance with reasonable robustness (stability). The out-of-the-box multinomial naive bayes has a mean cross-validation accuracy of approximately 62% with a validation accuracy of 64% (using a balanced 4k dataset).

Already, we can see that by labeling a small amount of data, we have achieved a satisfactory accuracy considering the baseline accuracy of a 4-class classification is just 25%. From now onward, we consider 64% to be the new baseline to improve upon.

4.3 Bidirectional Encoder Representations from Transformers (BERT)

In order to understand the mechanics of BERT, one needs to be familiar with a few fundamental components of BERT. The two most prominent features are self-attention and transformer.

4.3.1 Self-attention

Historically, deep learning models that are designed to solve NLP tasks typically rely on convolutional neural networks (CNN) and recurrent neural networks (RNN) to extract representation of feature. The former is capable of capturing local relation within a neighborhood, while the latter specializes in feature extraction from sequential data, e.g. a sentence. To allow smoother propagation of gradients through deep networks, researchers developed many variants of RNN such as the Long-Short Term Memory (LSTM) and the Gated Recurrent Unit (GRU). Nevertheless, one of the most pronounced limitations of RNNs is that they are by nature unidirectional, meaning information flows only forward. Additionally, it is challenging to parallelize propagation through RNN since computation must be done sequentially.

Self-attention, in contrast, is free of the issues above. Rather than processing the input sequentially, self-attention applies a mapping based upon trained weights (called **Query**, **Key** and **Value**) on every pair of input embedding and takes into account the locational information through positional embedding. Essentially, a self-attention layer maps the input sequence to a weighted representation by performing the following operation.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{d_k})V$$

The output representation is typically passed to another attention layer for further feature extraction. This architecture is characterized by the "attention weights, " a collection of normalized weights that govern the pair-wise mapping. A key virtue of the attention mechanism lies in its bidirectional nature, which gives rise to more flexible information flow in comparison to RNN.

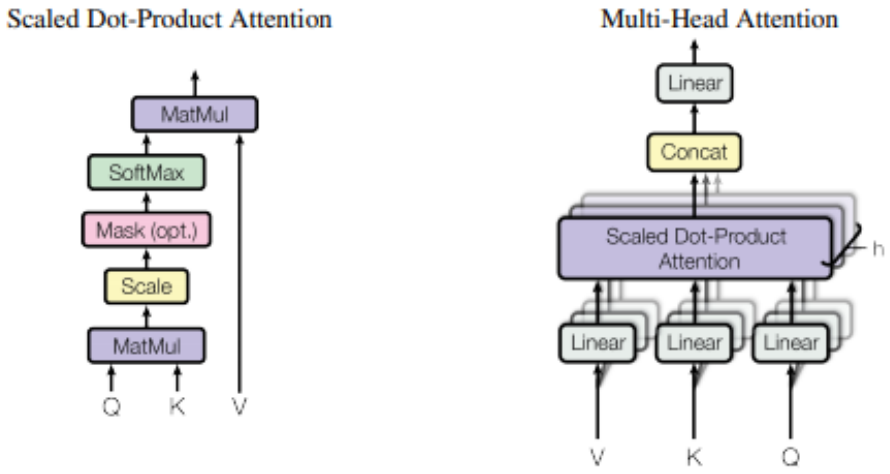

Figure 3: Illustration of attention

4.3.2 Transformer

Transformer is a deep learning framework that consists of stacked encoder and decoder blocks for solving various NLP tasks. It exclusively employs the **self-attention** mechanism for feature extraction (no CNN or RNN layer). Each encoder/decoder block consists of a multi-head attention layer followed by a feed-forward network (typically a fully-connected layer).

In Google's 2017 paper *Attention Is All You Need*, both the stacked encoder and the stacked decoder are composed of 6 identical blocks, as such arrangement achieves a desired balance between computational cost and complexity. Remarkably, empirical studies have shown that the transformer achieved superior performance over CNN and RNN-based architectures in varied downstream tasks.

4.3.3 BERT

The Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained, general-purpose deep learning architecture based upon the transformer described in the previous section. The most distinguishing characteristic is that BERT does not target at any particular downstream NLP problem. Rather, it endeavors to generalize to any potential task and has empirically stood out as the state-of-the-art NLP model. Since its debut in 2018, numerous variants have been crafted (e.g. Facebook's RoBERTa) and achieved excellent results on various benchmark NLP tasks. BERT's popularity can be attributed to its

1. flexibility, as users could fit the model into a larger framework that aims at a specific downstream task. In our case, topic modeling.
2. favorable performance given a small amount of data. In this project, since a very limited amount of labeled data is available, BERT's advantage in this respect proves to be indispensable.

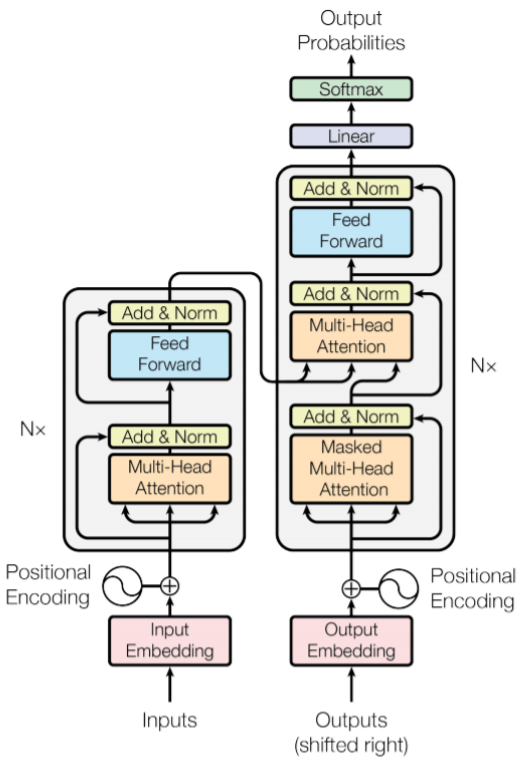

Figure 1: The Transformer - model architecture.

Figure 4: Transformer's architecture

4.4 Training Schemes

In this project, we primarily opted for two training paradigms to approach the topic modeling task: supervised learning and semi-supervised learning based upon BERT.

4.4.1 Supervised Learning

Supervised learning describes a training scheme that exclusively relies on labeled data to train a machine learning model. In the context of this project, we construct a deep neural network and train it using the labeled training data only.

In essence, our deep neural network classifier consists of two components: a feature extractor block and a classifier block.

Feature extractor block
For the feature extractor block, we employ a BERT structure with pre-trained weights that were trained on two sources of giant corpus (Wikipedia and BookCorpus). Note that a multitude of settings are available and, after thorough consideration, we decided to introduce the uncased base BERT, configured with 12 layers, 768 hidden size, 12 heads and 110M parameters.

The intuition here is that this setting features a relatively simple architecture and requires less computation, while it still manages to capture the semantic information sufficiently. Further, the uncased version was selected due to the fact that reviews, written carelessly without any grammatical guidelines, are often incorrectly capitalized.

It is worth noting that whether to fine-tune the weights in the BERT feature extractor during training is at our discretion. In reality, we experimented with both approaches and assessed the impact of fine-tuning empirically. Namely, with weights frozen or unfrozen. Conventionally, fine-tuning the

weights in BERT requires a small learning rate (the paper recommends 2e-5), otherwise serious overfitting would occur.

Classifier block
For the classifier block, we construct a feed forward network with a flexible number of layers.

As mentioned above, the BERT architecture was formulated as a feature extractor for NLP tasks in general. The output of BERT, in this context, is an array of shape sequence_length * hidden_size (768) given a single input. Hence, it is necessary to add an end block that serves our goal. Namely, to conduct text classification.

A common practice is to create a sequence of fully-connected networks with a varying number of layers. As no single rule governs the complexity of the network, two sensible options turn out to be a 1-layer classifier and a 3-layer classifier. For the latter, we employ the Rectified Linear Unit (ReLU) as the activation function and apply dropout (with probability 0.1) prior to every fully-connected layer for regularization purposes. The classifier block outputs a vector of length 4 (the number of desired classes). By applying a Softmax function to the raw output, we normalize the vector so that it shows the confidence associated with each class. A normalized output facilitates further analysis (e.g. precision-recall, ROC curve) and enhances the model's interpretability for both engineers and end users.

The diagram below offers a more illustrative view of the network structure.

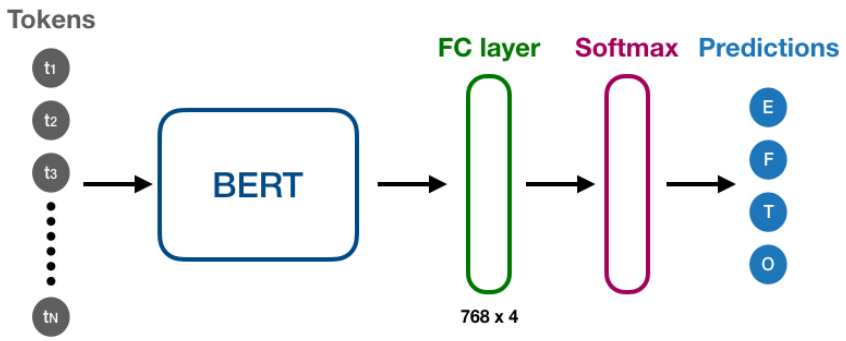

Figure 5: Architecture of BERT model with 1 classifier layer

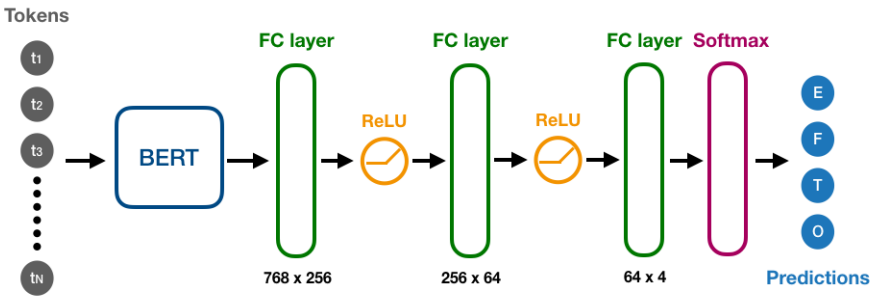

Figure 6: Architecture of BERT model with 3 classifier layers

4.4.2 Semi-supervised Learning

In recent years, semi-supervised learning has been an increasingly popular approach in place of purely supervised learning. Unlike supervised learning, which completely discards unlabeled data, semi-supervised learning makes efficient use of both labeled data and unlabeled data, as unlabeled data abound yet labeled data require much more effort to acquire. In our case, a liberal amount of unlabeled review texts are provided, but we have to manually label a tiny fraction of the reviews.

Training Accuracy on BERT Models				
	Supervised Frozen	Supervised Unfrozen	MLM Fine-tuned Frozen	MLM Fine-tuned Unfrozen
1-layer classifier	0.6855	0.9710	0.7451	0.8411
3-layer classifier	0.7256	0.8738	0.7699	0.8909

Table 4: Accuracies on training dataset

Broadly, semi-supervised learning consists of two stages.

First, unsupervised stage. Unlabeled data are used to allow the model to learn general features of the input data. A common architecture that caters to this need is an auto-encoder.

Second, supervised stage. Once the model is trained on unlabeled data, we fit the feature extractor into another (or augmented) framework that serves the end task and train it on labeled data.

BERT offers a graceful way, called Masked Language Modeling (MLM), to fine-tune weights on a new body of text. Briefly speaking, during fine-tuning, 15% of the tokens are corrupted at random (with 80% masked, 10% replaced with a random token and 10% unchanged). The model is then trained to predict the original tokens behind the corrupted tokens, which boils down to a classification problem. This approach has demonstrated effectiveness in various settings.

Essentially, the review texts are concatenated into a gigantic corpus and partitioned into equally sized chunks (512 tokens each). Next, weights are trained in the aforementioned manner chunk by chunk, thereby adapting to the idiosyncrasies in the POI review texts. After training 3 epochs, we reduced the validation perplexity from **20.9619** to **7.2945**, suggesting that the MLM effectively learned high-level features of the review texts.

Likewise, we conduct supervised learning on the labeled training data in accordance with the framework in the previous section. Nevertheless, the fine-tuned BERT feature extractor is used, rather than a pre-trained, out-of-the-box BERT.

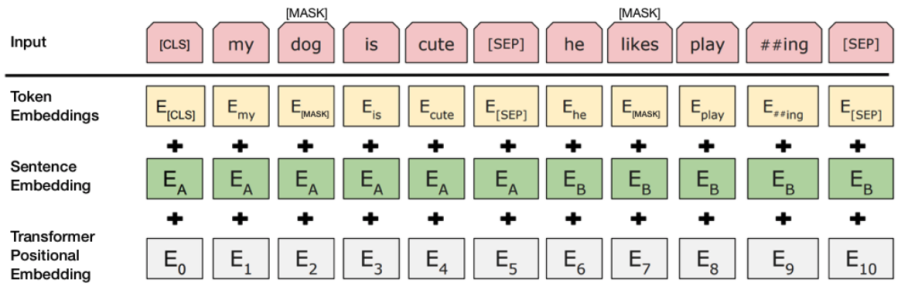

Figure 7: Illustration of fine-tuning BERT via Masked Language Modeling (MLM)

5 Evaluation

5.1 Overview

Comparing test accuracies between LDA/Guided LDA, conventional machine learning models and BERT, we find that BERT has the best performance both before and after tuning. BERT has a best validation accuracy of 0.7748 on our manually labeled dataset. This aligns well with our initial inclination due to the popularity and perceived advantage of deep learning models in NLP.

5.2 Ablation studies on BERT

Despite the excellent performance that an artificial neural network (ANN) is capable of achieving, one of its downsides lies in its obscurity. In other words, it is challenging to understand what actually goes on underneath the hood. As a result, people commonly regard it as a black box. Nevertheless,

Test Accuracy on BERT Models				
	Supervised Frozen	Supervised Unfrozen	MLM Fine-tuned Frozen	MLM Fine-tuned Unfrozen
1-layer classifier	0.6559	0.7079	0.7203	0.7673
3-layer classifier	0.7079	0.7475	0.7327	0.7748

Table 5: Accuracies on test dataset

until recently, there have been numerous attempts to dissect neural networks through a multitude of methods in hopes of shedding light on their mechanics.

5.2.1 Re-randomization of selected encoder block in BERT

Originally practiced in neural science, ablation studies have gained momentum over the past few years as a practical way to understand ANNs. They aim to identify components that contribute the most to the net performance of a neural network. In this project, we adopted a method called "re-randomization," proposed in Google's 2019 paper *Are All Layers Created Equal?* Briefly speaking, we filled the weights and biases of a hidden layer (or several hidden layers) with random values that are sampled from the original distribution of the trained parameters. Note that here normality is assumed. The rationale behind such practice is that, if a layer plays a critical role in producing an accurate prediction, then randomizing its parameters would severely impair the entire network and we expect a dramatic decline in accuracy. On the other hand, randomizing a layer of little importance would lead to a trivial drop in model performance.

Recall that BERT consists of 12 stacked encoders. In each experiment, we select one encoder and randomize all its weights and biases while keeping the other encoders intact. Then we generate predictions using this "corrupted model" on the test set and compute the accuracy. We repeat the experiment for 15 iterations and retain the average accuracy to account for randomness. Per the bar chart, our results exhibit that re-randomization invariably undermines the model, yet such effect is more pronounced in the shallower layers and dwindles as we go deeper. We speculate that, due to BERT's sequential nature, error travels and thus compounds over a longer path if a shallow layer is corrupted and vice versa.

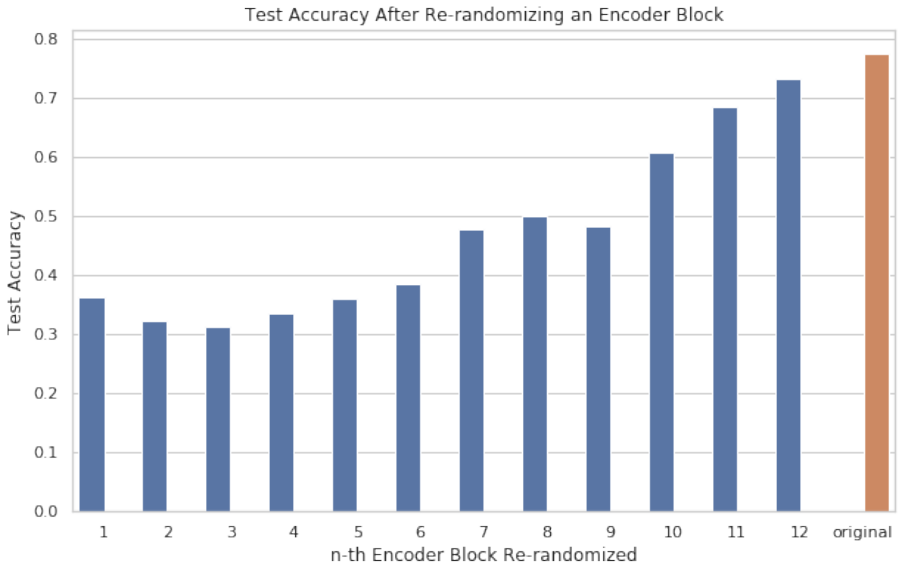

Figure 8: Test accuracy after a selected encoder block is re-randomized

5.2.2 Investigation into incorrect predictions

Additionally, an error analysis was conducted to investigate the causes of misclassification. We created a confusion matrix to offer extra insights. Notably, on the test set, the model exhibits an inclination to classify unseen sentences to the Tips class (29.4% of the predicted Tips are false). We therefore examined the associated misclassified observations closely. In simple terms, the pattern can be summarized as follows:

1. The sentence contains the word "you." Example: You could see sharks, all kinds of fish, turtles, etc. . . They have a Manatees, also called Sea Cows, witch are very fun to see because they are very peaceful, and life their life in slow motion. (True label: F)
2. The sentence conveys multiple layers of sentiment. Example: There are no portable toilet and the wait out in the cold and rain can easily be an hour so be prepared. (True label: F)

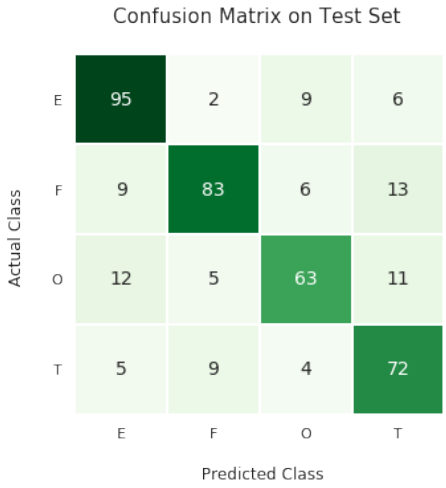

Figure 9: Confusion matrix on test dataset

6 Conclusion and Future Work

6.1 Conclusion

6.1.1 Overview

BERT has by far the best accuracy performance at **0.7748**, considering conventional model achieves 0.64 and a baseline accuracy would be 0.25 for a 4-class classification.

Overall, both our mentors from TripAdvisor and us are satisfied with the performance of the model and agree that the model can be deployed into production at the company. While the best-performing version of BERT takes anywhere from 1.5 hours to 3 hours to train depending on computational resources, it only runs on static, back-end data. Thus, we can be deployed and trained without affecting production.

6.1.2 BERT has high confidence in its predictions

Additionally, BERT is also confident when it comes to its predictions. Using a softmax function, BERT outputs a probability distribution as shown in Figure 10 that gives a high confidence score to one class. This is not observed in the conventional supervised models, where the distribution is more even. In fact, because of the generally low confidence score of the conventional supervised models, their accuracy decreases as one increases the threshold for a particular class of prediction.

Such distribution in confidence level makes a good business case where we can highlight the most confidently predicted results to the liking of a visitor to TripAdvisor's website, knowing that it has over 95% probability of being the right label class.

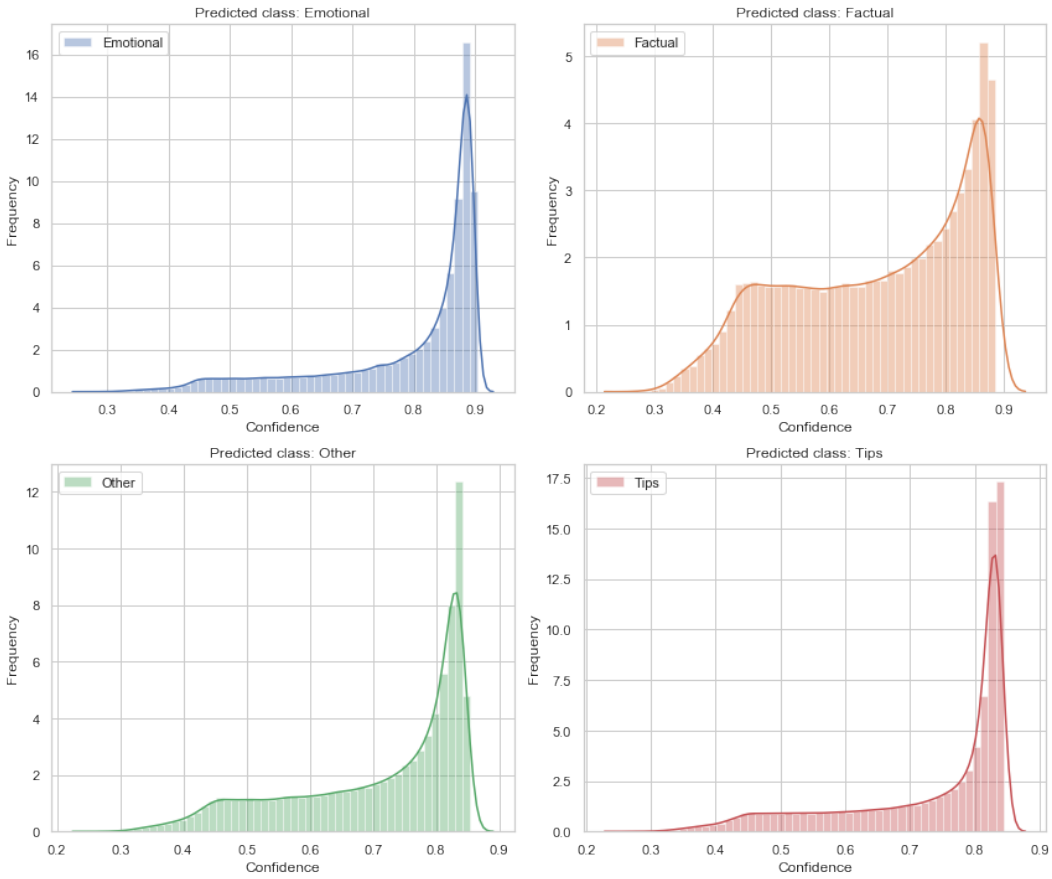

Figure 10: Distribution of Prediction Confidence among 4 Classes (BERT)

6.2 Future work

6.2.1 Ensure good representation of output with regard to original data set

The business application of our model and output is versatile. A suggested use case is to produce the top-5 most confidently predicted tips and factual reviews for a point-of-interest and highlight them on the website. Based on initial observation, however, we have noticed that many of the most confidently predicted reviews are well-written and grammatically correct. This is somewhat surprising because they defy the popular perception of the quality online reviews. While only minimal processing occurred prior to fitting the model to data, namely to parse the review paragraphs into sentences, and the corpus was in no way beautified, more investigation may be needed to ensure good representation of the highlighted results in relation to their original format.

Likewise, we noticed based on empirical observation that the top-5 predicted emotional reviews are all of positive sentiment for the point-of-interest locations we looked at. While the prediction is sound and useful to any users online, more investigation may therefore be warranted to see if this is due to the pattern of the underlying data set, or if there is bias in the model towards positive sentiment. Specifically, it would be ill-advised to highlight only positive reviews on the website when the vast majority of emotional reviews are negative for a particular tourist attractions. In such a case, there may be a trade-off between high confidence, and therefore correctness, and staying true to the prevailing sentiment of the tourist attraction.

6.2.2 Continuous training and evaluation

Our best model was only trained on 3,000 rows on manually labeled reviews and additional labeled non-reviews. We would suggest that TripAdvisor continue to re-train the model on new user reviews

to keep the model up-to-date. An obvious way to obtain more labeled data would be to retain services such as Amazon Mechanical Turk, in order to label more user reviews. This necessarily diversifies the viewpoints going into the labeling process by involving more people, helping the labeled data to become more representative of what a regular user of the website might think and not restricting the perceived labels to only three people. Additional labeled data that is not from user reviews may also be obtained, which conforms to the four classes that we are interested in.

Another potential approach that does not involve labeling new user reviews may be to use the model to output results that it is the most confident about, and use those results as training data together with any labeled or old data that the company might have. For example, we can predict the top-5 most confident reviews for each class for each POI, which would created additional $5 \times 4 \times$ (number of POI) data points to train with. Because the confidence is high for the predicted results, we can be comfortable using them as "pseudo" ground truth.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[4] Pritchard, J.K. & Stephens, M. & Donnelly, P. (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* **155**(2):945-959

[5] Jagarlamudi, J. & Daume III, H. & Udupa, R. (2012) Incorporating Lexical Priors into Topic Models. *Association for Computational Linguistics Anthology* **E12-1021** pp. 204–213.