New Evaluation Methodology for Qualitatively Comparing Classification Models

Ahmad Aljanaideh
Bentley University
aaljanaideh@bentley.edu

Abstract

Text Classification is one of the most common tasks in Natural Language Processing. When proposing new classification models, practitioners select a sample of items the proposed model classified correctly while the baseline did not, and then try to observe patterns across those items to understand the proposed model's strengths. However, this approach is not comprehensive and requires the effort of observing patterns across text items. In this work, we propose a new evaluation methodology for performing qualitative assessment over multiple classification models. The proposed methodology is driven to discover clusters of text items where each cluster's items 1) exhibit a linguistic pattern and 2) the proposed model significantly outperforms the baseline when classifying such items. This helps practitioners in learning what their proposed model is powerful at capturing in comparison with the baseline model without having to perform this process manually. We use a fine-tuned BERT and Logistic Regression as the two models to compare with Sentiment Analysis as the downstream task. We show how our proposed evaluation methodology discovers various clusters of text items which BERT classifies significantly more accurately than the Logistic Regression baseline, thus providing insight into what BERT is powerful at capturing.

Keywords: Evaluation Methodology, Sentiment Analysis, Word Embeddings

1. Introduction

Text Classification is a very common and important Natural Language Processing task with various applications such as Sentiment Analysis and Spam Detection. Practitioners build new classification models and perform qualitative and quantitative analysis on how those models compare to baseline models. Those kinds of analyses are important to assess the newly proposed classifiers and the nuance they capture more richly than the baseline models.

Quantitative analysis is often performed by comparing the accuracy of the proposed classification model to that of the baseline model. On the other hand, qualitative analysis is performed by selecting validation items that the proposed model classified more accurately than the baseline, and then trying to observe linguistic patterns across validation text items. This is done to obtain insight into what the proposed model is powerful at capturing. However, this approach is not comprehensive as practitioners tend to select only a few items. Moreover, it requires examining patterns across such items which can take time and effort since most datasets contain thousands of text items.

Various techniques have been proposed for explaining a model's decisions, such as the attention mechanism (Bahdanau et al., 2014), Saliency Maps (Li et al., 2015), LIME (Mishra et al., 2017), SHAP (Lundberg and Lee, 2017), BETA (Lakkaraju et al., 2017) and BERT Clustering (Aljanaideh, 2022). Moreover, previous work investigated the inner workings of neural models and what they capture (i.a. Tenney et al., 2019; Rogers et al., 2021; Ebrahimi et al., 2017; Rogers et al., 2021; Clark et al., 2019; Niven and Kao, 2019). However, all of those techniques focus on investigating and explaining decisions made by a stand-alone model and thus do not provide insight into how decisions by two models differ across different text items.

In this work, we propose a new evaluation methodology for qualitatively comparing the performance of classification models. The proposed approach focuses on automatically discovering text patterns correlated with items that a proposed model predicts more accurately than a baseline model. The proposed approach is inspired by Aljanaideh et al. (2020)'s approach of clustering contextualized BERT embeddings (Devlin et al., 2018). The approach is based on splitting occurrences of a word across different text items into different clusters using the embeddings of the word and the predictions of the models being compared.

We apply our evaluation methodology on a fined-tuned BERT and Logistic Regression models with Sentiment Analysis as the downstream task. We chose those models and this task since they are widely popular and to obtain insight into when looking at the context of words is useful. We use short reviews sampled from the Yelp dataset[1]. We perform cluster analysis on the discovered patterns to understand the kinds of items BERT predicts more accurately than the logistic regression baseline.

Results show our approach discovers interpretable clusters of items which BERT classifies significantly

[1]Available at https://www.yelp.com/dataset

LREC-COLING 2024, pages 12187–12192
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

more accurately than logistic regression. Examples include items in which mixed sentiment is expressed (e.g. *The food was good, but the service was bad*). We conclude by encouraging practitioners to work towards developing automatic qualitative assessment techniques to obtain insight into what proposed classifiers offer.
In Section 2, we describe our proposed evaluation methodology. In Section 3, we describe the classification models and dataset we use to assess the proposed evaluation methodology. We show the results in Section 4 and finally conclude in Section 5.

2. Proposed Evaluation Methodology

We propose an evaluation methodology for automatically discovering the kinds of text items a classification model outperforms a baseline model at. The proposed methodology relies on three steps. First, we obtain predictions from each of the two models on validation text items using 5-fold cross validation. Second, we pass those text items along with the predictions to a clustering model. The clustering model is driven to discover clusters of items which are linguistically similar and where the proposed model significantly outperforms the baseline. Third, we rank the discovered clusters based on the percentage increase in accuracy when using the proposed the model vs. the baseline.

2.1. Obtaining Predictions

The first step is obtaining predictions from the two classification models. The two models are trained separately for a downstream task (e.g. Sentiment Analysis) using 5-fold cross validation. The outcome of this step is a prediction from each model for every text item in the dataset.

2.2. Pattern Discovery

In the second step, we use the text items and the predictions of each two models (obtained in the first step) to discover the kinds of text items the proposed model significantly outperforms the baseline. We leverage Aljanaideh et al. (2020)'s clustering approach for this step and modify it for our purpose. Next, we describe Aljanaideh et al. (2020)'s clustering model, and then we illustrate how we modified it to fit our evaluation methodology.

2.2.1. BERT Clustering Model (Aljanaideh et al., 2020)

Aljanaideh et al. (2020) developed a model that discovers fine-grained context patterns of words from labeled text items. To achieve this, the model leverages pre-trained contextualized BERT embeddings (Devlin et al., 2018). The model takes as input multiple embeddings of the same word and splits them into different clusters using the embeddings of the word and the labels of the items the word appeared in. The approach is a decision-tree where at each step, the embeddings are split into two clusters recursively. Given that there is more than one way to split a number of points into two clusters, the split which achieves in the maximum information gain is the one which is selected. The splitting is terminated when a cluster of 100% purity is achieved or when a depth threshold is reached. The depth threshold is determined using the common logarithm of the number of embeddings which is equivalent to the frequency of the word in the dataset. The clustering algorithm is applied on every word in the training set. The result is a set of clusters for every word where each cluster contains items which are linguistically similar and where the items are dominantly positive or negative.
The model was applied for the task of detecting politeness in requests. The authors demonstrated that it is able to automatically discover interpretable patterns correlated with (im)polite language. For example, they showed that the use of the word *please* with a direct tone (e.g. *Would you please stop?*) is correlated with impolite requests while requests which start with a greeting and then use the word *please* in a request sentence (e.g. *Hello! Could you please help me?*) are correlated with polite requests.

2.2.2. BERT Clustering in the Proposed Evaluation Methodology

We use an approach similar to Aljanaideh et al. (2020)'s approach for our evaluation methodology. Similar to their approach, we apply a decision-tree model to cluster the pre-trained BERT embeddings of every word in the dataset. We also look for the split which achieves the maximum information gain and recursively repeats the splitting. However, we modify two components of their model to fit our propose. While their model used the labels of the downstream task to perform the pattern discovery, we use a different labeling scheme which fits our goal. Specifically, we label an item in the dataset as positive if the proposed model predicted it correctly while the baseline did not. Otherwise, the item is labeled as negative. This labelling fits our evaluation methodology as the goal is to discover items at which a classifier significantly outperforms another classifier. Moreover, Aljanaideh et al. (2020) decision-tree terminates when a cluster of 100% purity is achieved (calculated using the labels of the downstream task) or when a maximum depth threshold is reached. In our case, the clustering is terminated when the two models have the same predictions for all items in the cluster, or when the depth threshold is reached.

2.3. Cluster Ranking

The third step is ranking the clusters obtained in the previous step. For each cluster of items, we calcu-

late the percentage increase in accuracy on those items when using the proposed model vs. when using the baseline model. The clusters are ranked using this value such that top-ranked clusters highlight items with a significant percentage increase in accuracy when using the proposed model. This helps in learning what the proposed models captures significantly more accurately than the baseline.

3. Application

In this section, we describe the dataset and classification models we use to assess our evaluation methodology.

3.1. Dataset

We use the Yelp Dataset. This dataset contains customer reviews for various businesses. Each review is labeled with a star rating in the range [1, 5]. We sample 2,000 reviews. We follow previous work by labelling reviews of 1 or 2 stars as negative and reviews of 4 or 5 starts as positive. We only consider reviews of less than 50 words to facilitate interpretation when performing the cluster analysis. The data is balanced across the two classes. We split the data into 80% training and 20% testing. The training portion is used to perform the cross-validation and cluster analysis. The test portion is simply used to evaluate the classification models on the Sentiment Analysis task.

3.2. Classification Models

We use a fine-tuned BERT and a Logistic Regression as the two classification models to compare. The logistic regression model is trained using unigram features while the fine-tuned BERT model is used as described in Devlin et al. (2018). We chose those two models since are they are widely used and also to obtain insight into when context of words is useful as BERT considers the context of words while the logistic regression unigram-based model does not. Both models are trained using 5-fold cross-validation. For each of the two models, we obtain a prediction for each item. We then apply the clustering model described in Section 2.2.2 to obtain clusters of items. The clusters are then ranked such that clusters that contain items where BERT significantly outperforms the Logistic Regression baseline are ranked at the top. Next, we perform cluster analysis on top-ranked clusters.

4. Results

The test accuracy of the fine-tuned BERT model on the Sentiment Analysis task is 95.7% while the accuracy of the logistic regression model is 91%. Next, we perform cluster analysis on the training portion to assess what the BERT model excels at the most in comparison with the logistic regression model. We next perform analysis over the clusters discovered by our model.

Overall, our model discovered 7800 clusters from 5379 words. Each word is associated with multiple clusters. Table 1 shows a sample of the top-ranked clusters. For each cluster, we show the corresponding word, its size in terms of number of items, the accuracy for each of the two models on those items, the percentage improvement in classification accuracy when using the fine-tuned BERT vs. Logistic Regression model, the pattern observed over such items, example items and the binary sentiment label of each item.

The first cluster corresponds to the question mark. BERT outperformed the baseline with a 500% improvement on items in this cluster. The cluster contains items where customers expressed an extremely negative sentiment using repeated punctuation (e.g. *How does this guy stay in business??*).

The second cluster corresponds to the word *best*. The cluster contains reviews in which the customer used this word but expressed negative sentiment (e.g. *wasn't the best*). BERT outperformed the baseline by 100% on items in this cluster.

The third cluster corresponds to using the word *better* but in reviews where the customer is dissatisfied with their experience (e.g. *better food can be found somewhere else*). BERT outperformed the baseline by 93%.

The fourth cluster corresponds to the negation word *not*. The cluster contains reviews in which the customer negated a positive word (e.g. *this pops up as the best place available. It's definitely not*). BERT outperformed the baseline by 50%.

The fifth cluster corresponds to the word *great*. Reviews in this cluster used this word to express mixed sentiment (e.g. *Food and beer was great. Very dissatisfied with service.*). BERT outperforms the baseline on such items by 67%.

The sixth cluster shown in the table corresponds to the word *back*. In items in this cluster, reviews expressed their dissatisfaction and how they do not intend on going back to the place they are reviewing (e.g. *I won't be going back*). BERT outperforms the baseline by 30% on such items.

The last cluster was obtained by clustering the CLS BERT token which encodes the entire text item. The cluster contains relativity short reviews in which customers expressed positive views but without using strong sentiment words (e.g. *love*).

Overall, BERT is significantly better than logistic regression at handling mixed sentiment and presence of both positive and negative words. However, depending the pattern, the level of improvement when using BERT varies. This clustering provides insight to the level of improvement depending on the pattern.

Table 1: Clusters discovered with our model. For each cluster we show the corresponding word, its size, the logistic regression model accuracy (LRA), BERT model accuracy (BA), the percentage improvement in accuracy when using BERT vs. the baseline (%Δ), two example reviews, and the sentiment labels of those reviews.

Cluster	Size	LRA	BA	%Δ	Pattern	Examples	Label
?	12	17%	100%	+488%	Emphasizing with Repeated Punctuation	I'm glad this place isn't closer. Srsly, Peanut Butter & Chocolate, where did you come from?!	+
						Employees never have grille items!!!! Always slacking at the easiest job ever**!**!!! How does this guy stay in business**?**? [...]	-
best	16	50%	100%	100%	Using the word best but in a negative/mixed sentiment	My husband and I ate lunch here a month ago and the experience wasn't the **best**.[...]	-
						Pretty average take-out chinese. Not bad, not the **best**, but id definitely eat there again [...]	-
better	11	55%	100%	83%	you can find better places	Service was good. Ambience good. **better** food can be had elsewhere. Priced too high for taste, quality and quantity.	+
						For the price of the room we should of had **better** service and more perks....	-
not	41	59%	88%	50%	Negating positive words	This place really isn't that good. The atmosphere is sweet, but the food is mediocre at best. I'm always disappointed when I type in middle eastern and this pops up as the best available, it's definitely **not.**	-
						he food was tasty service **not** so good two of us got our plates then the third person	-
great	11	55%	91%	65%	Using the word great but expressing mixed sentiment	Food and beer was **great**. Very dissatisfied with the service[...]	+
						We went back for the same reason as in 2011 - we thought this was a different Italian Restaurant. Again, we were pleasantly surprised - the food and the service was **great**!	+
back	27	74%	96%	+30	Customers indicating they don't intend on going back to the business they're reviewing	They have changed the insol%es I Wont be going **back** there.	-
						Their sushi burrito was nowhere near as good ... I'll never go **back**.	-
CLS	19	68%	100%	47%	Relatively short positive reviews that lack sentiment words	I'm glad this place isn't closer. Srsly, Peanut Butter & Chocolate, where did you come from?!	+
						Don't worry when we make this mac and cheese business pop off we'll give u the word.	+

5. Conclusion

In this work, we proposed a new evaluation methodology for comparing classification models. The proposed methodology combines qualitative and quantitative analysis. The proposed method relies on recent embedding clustering techniques to discover items where a proposed model significantly outperforms a baseline model. We applied the proposed method on the task of sentiment analysis. We performed analysis on discovered patterns and showed the kinds of items a fine-tuned BERT outperforms a logistic regression model at.

In the future, we plan to expand this method to more Natural Language Processing tasks and models. Specifically, we plan to examine this method across a wide-array of classification models including those based on Large Language Models (LLMs). This will help in understanding capabilities of LLM-based classifiers. We also plan to expand this method to seq2seq tasks such as machine translation in order to understand the strengths or recent technqiues.

6. Bibliographical References

Ahmad Aljanaideh. 2022. *Leveraging Word Embeddings to Enrich Linguistics and Natural Language Understanding*. The Ohio State University.

Ahmad Aljanaideh, Eric Fosler-Lussier, and Marie-Catherine de Marneffe. 2020. Contextualized Embeddings for Enriching Linguistic Analyses on Politeness. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2181–2190.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does Bert Look at? an Analysis of Bert's Attention. *Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP at ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. 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*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box Adversarial Examples for Text Classification. *Association for Computational Linguistics*.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and Understanding Neural Models in Nlp. *North American Chapter of the Association for Computational Linguistics*.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. *Advances in neural information processing systems*, 30.

Saumitra Mishra, Bob L Sturm, and Simon Dixon. 2017. Local Interpretable Model-agnostic Explanations for Music Content Analysis. In *ISMIR*, volume 53, pages 537–543.

Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A Primer in Bertology: What we Know About How Bert Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. *ICLR*.