Sentiment Analysis of Amazon Reviews

Yu-Chi Tsao

yctsao@stanford.edu

Abstract

Substantial progress has been made on the task of amazon review prediction using machine learning and deep learning model, such as SVM, LSTMs, and GRUs. There has been research found out introducing some useful information by combination of customer review and produce embedding (Shrestha and Nasoz, 2019) or including user-written summaries of product reviews could improve the model prediction Yang et al. (2019). Hence, we would like to investigate how different feature representations going to affect the model prediction. The preliminary results show that the model performance for using BERT embedding for amazon review data is the best compared with bags of word or GloVe representations.

1 Introduction

Sentiment analysis interprets and classifies positive, neutral, and negative attitudes within text data. The applications of sentiment analysis include identifying customer satisfaction levels, understanding users' opinions or attitudes toward a topic, etc. Furthermore, sentiment analysis has been proven to be a valuable technique for other applications, like recommendation systems. Because of its wide range of applications, it has attracted lots of research interests in recent years.

Previous approach focused on amazon review classification has shown that introducing some useful information by combination of customer review and produce embedding, the accuracy can be improved from 81.29% to 81.82% (Shrestha and Nasoz, 2019). Another example of incorporating more useful information for sentiment analysis tasks to include user-written summaries of product reviews. Since a summary can be highly indicative of the polarity of a given text, it may be worth including this information for sentiment classifications. Yang et al. (2019) investigated a joint encoder, which can capture the interactions between review and summary information and evaluated their model on the SNAP (Stanford Network Analysis Project) Amazon review datasets. They showed a 4.8% improvement compared to the previous best method. In addition, another aspect to improve the model performance is to deal with the problem of class imbalance. The amazon review data is highly imbalanced and are skewed towards 4 and 5 stars. (Heidi et al. 2018) Hence, Mukherjee et al. (2019) used random oversampling method by oversampling the minority class (rating 1, 2, 3) to reduce the class imbalance. The minority class have been repeated multiple times to increase their occurrences. Through this method, Mukherjee et al. (2019) demonstrated that the accuracy for test set could be improved from 68.15% to 80.16% by using LSTM. Hence, we think having high model performance is highly related to the design of feature representations.

Our goal is studying different feature representation through introducing some useful information, testing different encoding method, and approaches to solve imbalanced dataset for predicting positive and negative reviews in amazon review data. Here we analyzed the preliminary results on testing different feature representations method through GloVe and BERT.

2 Related Work

2.1 Machine learning and lexicon-based approaches

Machine learning method using different aspects of text as sources of features have been proposed in the literature. Tomas et al. (2017) compared four most popular machine learning models: Naïve Bayes, Support Vector Machine (SVM), Decision Trees and Random Forest by using Amazon unlocked mobile phone review. They demonstrated that SVM could outperform regarding to accuracy, precision, recall and F1 score (F1 score = 89). Random forest can also achieve better results, just slightly worse than the SVM. But Naïve Bayes and Decision Trees perform similar and worse than SVM and random forest. The author stated the reason to have worse performance may be that these two algorithms have lower complexity when compared to random forest and SVM. [1]

Heidi et al. (2018) compared different machine learning model and lexicon-based approaches (VADER, Pattern, and SentiWordNet) by using amazon dataset with 44 product categories. The lexicon-based approaches associate words to their sentiment orientation represented for example by positive and negative scores. They found out that machine learning algorithm outperform the lexicon-based techniques and LR algorithms outperforms the SVM and Gradient Boost algorithm with the highest accuracy, precision, recall, and F1 score (F1 = 94). They also found the model performed better in classifying positive classes compared to negative classes, which they attributed the reason is due to the imbalanced dataset. (Heidi et al. 2018)

Aside from different machine learning model has been tested in the literature, using ensemble machine learning algorithm has been presented by Sadhasivam et al. (2019) The author stated that ensemble method generally provides better accuracy as it combines the algorithms that help in functioning effectively. Through using Naïve Bayes and SVM classifiers and then majority voting to choose using which model to classify the data, the accuracy could be improved from 37 % to 78 % by comparing just using Naïve Bayes model. The accuracy can be improved from 33 % to 73 % by comparing just using SVM. (Sadhasivam et al. 2019)

Hamouda et al. (2011) used a hybrid way of building machine learning based senti-word lexicon for sentiment analysis. They chose 25000 reviews randomly from amazon reviews, did the data preprocessing, and used SVM to train based on n-grams and TF-IDF. Then the generated model from this training contained bag-of-words with positive or negative weight values. Used those bag-of-words as a MLBSL and applied this lexicon on two corpora. First corpus is the amazon corpus, where they chose 4000 reviews (2000 positive and 2000 negatives). The second one used the movie corpus. Term counting was then applied and an accuracy of 71.75 % is achieved. Compared with other literatures with the published lexicon, their method perform better than others by using sentiwordnet-term counting (65.85%), term counting (69.35%) and term counting from combined lexicon and valence shifters (67.80%). (Hamouda et al. 2011)

2.2 Deep learning approaches

Deep learning for sentiment analysis has attracted great attentions in recent years. They have been shown to outperform traditional machine learning approaches, such as SVM and logistic regression (LR). Among the various deep learning models, recurrent neural network (RNN) has shown great promise in many natural language tasks and is a very popular model because it has the ability to preserve sequential information in the dataset. In a traditional neural network, all inputs are considered to be independent of each other. However, for many natural language tasks, this is usually not the case because of the dynamic behavior of human language.

Although RNN exhibits several advantages over other models, it is much difficult to train. Depending on the activation functions used in RNN, Jozefowicz et al. (2015) showed that the gradient is likely to vanish or explode during back propagation. In addition, if a sequence is very long, it is usually difficult for RNN to carry the information and pass it on. This would also cause some problems when the dataset is composed of long paragraphs of text. To overcome these challenges, long short-term memory (LSTM) and gated recurrent units (GRU) have been developed. (Hochreiter and Schmidhuber, 1997; Cho et al., 2014)

Long short-term memory (LSTM): LSTM, developed by Hochreiter and Schmidhuber (1997), consists of a cell (the memory part of a LSTM unit), an input gate, an output gate and a forget gate (the regulation part that controls the information flow into and out of the cell). Hence, LSTM are able to carry long-term memories, which are useful for several natural language tasks, such as handwriting recognition, machine translation and of course sentiment analysis.

Gated recurrent units (GRU): Similar to LSTM, GRU, which is introduced by Cho et al. (2014), reduces the gate networks to two with the output gate being removed. Although studies have shown that LSTM usually outperform GRU, GRU is computationally more efficient because of its less complex structure. Hence, more and more studies have been focusing on applying GRU in natural language tasks.

As discussed above, because RNN models possess many benefits in dealing with sequential data, several variants of RNN are being developed. Sachin at al. (2020) compared several baseline models using LSTM, GRU, bidirectional LSTM (Bi-LSTM) and bidirectional GRU (Bi-GRU) on an Amazon review dataset using GloVe word embeddings. They show that Bi-GRU model exhibits the best accuracy, although it took much longer time to train a bi-directional model. (Sachin et al., 2020)

In addition to developing new deep learning models, researchers are searching for more efficient ways to represent the text data in vector spaces and to capture more information that may be useful for model training. Nishit Shrestha et al. developed a model that incorporated both semantic relationship of Amazon product review and the product information. They argued that a product that receives a positive review from a customer is likely to also get positive reviews from other customers. Hence, including the product information may help. In their work, paragraph vectors were used to convert Amazon product reviews into fixed-length feature vectors. Then these feature vectors were used to train an RNN with GRU to get product embeddings. With the combination of customer review embeddings and product embeddings, they showed an improvement in accuracy from 81.29% to 81.82%. (Shrestha and Nasoz, 2019)

Another example of incorporating more useful information for sentiment analysis tasks is to include user-written summaries of product reviews if they are available. Since a summary can be highly indicative of the polarity of a given text, it may be worth including this information for sentiment classifications. Yang et al. (2019) investigated a joint encoder, which can capture the interactions between review and summary information, and evaluated their model on the SNAP (Stanford Network Analysis Project) Amazon review datasets. They showed a 4.8% improvement compared to the previous best method.

Another aspect to improve the model performance is to deal with the problem of class imbalance. The amazon review data is highly imbalanced and are skewed towards 4 and 5 stars. (Heidi et al. 2018) Hence, Mukherjee et al. (2019) used random oversampling method by oversampling the minority class (rating 1, 2, 3) to reduce the class imbalance. The minority class have been repeated multiple times to increase their occurrences. Through this method, Mukherjee et al. (2019) demonstrated that the accuracy for test set could be improved from 68.15% to 80.16% by using LSTM.

3 Model and Feature Representations

The baseline models we used is logistic regression with bag of words representations. We then changed the bag of words representation to GloVe or BERT and compared the model performance with the baseline model. For logistic regression, we did the parameter search when we used different feature representations.

4 Data

Amazon camera reviews are chosen to be the dataset to study for sentiment analysis. Each row of the dataset consists of several information, including customer_id, product_id, star_rating, review_body, etc. A sample dataset is shown in figure 1. Each product review is rated from 1 star to 5 stars. There are 4083 reviews where the star is below or equal to 3 and 13925 reviews where the star is above 3. Here we treat 1-3 star reviews to be negative reviews and 4-5 stars to be positive reviews. Hence, we have imbalanced dataset with

13925 positives and 4083 negative reviews. We compared different feature representation by using 0.001% of dataset and ran at our laptop. But eventually we would like to run all the datasets in the AWS. And the dataset is divided into training (70%) and test sets (30%).

We compared different feature representation by using 0.001% of dataset and ran at our laptop. But eventually we would like to run all the datasets in the AWS.

market place	customer _id	review _id	product _id	product _parent	product _title	star _rating	helpful _votes	total _votes	vine	verified _purchase	review _headline	review _body	review _date
US	12491786	R1C0EVYBV2Z18Z	B000U92DLA	8E+08	Kastar Camera & Camcorder Battery Home Travel ...	5	6	6	N	Y	A Must For On The Go	Purchased the charger from Nextop2100, arrived...	11/22/2008
US	37280009	R3Q837V78NS6B9	B0008GCYNW	6E+08	Olympus Camedia C755 4MP Digital Camera with 1...	3	15	17	N	Y	The Good and the Not-so Good	I was attracted to this camera for three reaso...	9/8/2005
US	16532896	R22CZCJH5YNX1O	B003CZ7L84	9E+08	Intova CP-9 Compact Digital Camera with 130 fe...	4	1	1	N	Y	Great for what we paid!	We took it with us to Cancun. My husband took ...	11/2/2011

Figure 1. A sample dataset

5 Metric

Due to the imbalanced datasets where the class distribution is not uniform among the classes, using accuracy as metric is not ideal since it fails to control for size imbalances in the classes and just assume all the dataset with equal distribution. Precision and recall are the standard metrics for binary classification. To see the overall effect on the precision and recall for positive and negative review, we used F1 score which combines precision and recall and give equal weight to precision and recall. Because the F score does not have normalization of the size of the data sets with K category, Individual scores for a category may be misleading about the overall performance of the system. Hence, we also used macro-average F1 score, which is the mean of F1 score for each category (positive and negative reviews). The reason to use macro-average F1 score is that compared with micro-average or weighted F1 score, it gives equal weight to each category. Considered with the datasets are imbalanced, macro-average F1 score will be a better metric to compare the overall model performance results and get full understanding on the model results.

6 Results and Discussion

GloVe embedding method constructs an explicit word-context or word co-occurrence matrix using statistics across the whole text corpus. The result is a learning model that may result in generally better word embeddings. Bert (Bidirectional encoder representations for transformers) has recently been considered a powerful encoding method by applying the bidirectional training of transformer, a current attention model, to language modelling (Devlin et al. 2018). Hence, here we used these two different embedding methods and compared their macro-average F1 score with the baseline model's macro-average F1 score. When we changed the embedding from bag of words to GloVe, we found both the training and test macro-average F1 increased and overfitting is reduced (Table 1). Training F1 increased from 0.97 to 1 and test F1 increased from 0.65 to 0.81. Furthermore, when we changed to BERT, the test macro-average F1 could even further increase to 0.87, which is the highest among these three different feature representations. Hence, we demonstrated using powerful embedding method (BERT) could enhance the model predictions for amazon reviews.

	Baseline model		
Feature representation	Bag of words	GloVe	BERT
model	Logistic regression	Logistic regression	Logistic regression
Training (macro-average F1)	0.97	1	1
Test (macro-average F1)	0.65	0.81	0.87

Table1. Macro-average F1 results with three different feature representations.

7 Conclusion and Future work

The preliminary results show that the model performance for using BERT embedding for amazon review data is the best compared with bags of word or GloVe representations. This indicates feature representations are playing important role for improving model prediction. One limitation with BERT for amazon datasets is that we found reviews in amazon data have larger than 512 tokens. Hence, we need to truncate the review length. Another improved way is perhaps we could use RoBERTa. In addition, maybe we can try to introduce more information. Currently we used the review content as feature. The datasets actually contain information of product review summary (the abstract concept of the entire review), review, helpful votes (number of people who think this review is helpful), and total votes. We hypothesized combining that information together may help the model result since helpful votes may be a good indication whether the review content is fake reviews or not. Or an ensemble approach through extracting feature pertaining to sarcasm, humour, hate speech and feed them to the model have shown by Badlan (2019) to lead to a better empirical performance for yelp review sentiment classification than a model that predicts sentiment alone. Hence, maybe we could extract features like fake reviews to investigate the effect on the model prediction. Finally, we would like to combine with other amazon review data (books, cell phone) and run in AWS to see if our results could be generalizable.

8 Acknowledgments

All authors contributed equally to this work. Yu-chi Tsao is responsible for implementing BERT and GloVe, training the model, and providing the corresponding analysis. Tzu-ling Liu is responsible for implementing baseline models.

References

Tomas Pranckevicius and Virginijus Marcinkevičius. 2017. *Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification.* Baltic Journal of Modern Computing.

Heidi Nguyen, Aravind Veluchamy, Mamadou Diop, and Rashed Iqbal. 2018. *Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches.* SMU Data Science Review: Vol. 1: No. 4, Article 7.

Jayakumar Sadhasivam and Ramesh Babu. 2019. *Sentiment Analysis of Amazon Products Using Ensemble Machine Learning Algorithm. International Journal of Mathematical.* Engineering and Management Sciences. 4. 508-520.

Dr Hamouda, Mahmoud Marei, and Mohamed Rohaim. 2011. *Building Machine Learning Based Senti-word Lexicon for Sentiment Analysis.* Journal of Advances in Information Technology. 2. 10.4304/jait.2.4.199-203.

Rafal Jozefowicz, Wojciech Zaremba and Ilya Sutskever. 2015. *An empirical exploration of recurrent network architectures.* In: International conference on machine learning; 2015. p. 2342–50.

Sepp Hochreiter and Jürgen Schmidhuber 1997. *Long short-term memory.* Neural Comput. 9(8):1735–80.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. *Learning phrase representations using RNN encoder–decoder for statistical machine translation.* In: Proceedings of the 2014 conference on empirical

Sharat Sachin, Abha Tripathi, Navya Mahajan1, Shivani Aggarwal, and Preeti Nagrath. 2020. *Sentiment Analysis Using Gated Recurrent Neural Networks.* SN Computer Science (2020) 1:74

Nishit Shrestha and Fatma Nasoz. 2019. *Deep Learning Sentiment Analysis of Amazon.com Reviews and Ratings.* International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.8, No.1, February 2019

Sen Yang, Leyang Cui, and Yue Zhang. 2019. *Exploring Hierarchical Interaction Between Review and Summary for Better Sentiment Analysis.*

Anirban Mukherjee, Sabyasachi Mukhopadhyay, Prasanta K. Panigrahi, Saptarsi Goswami. 2019. *Utilization of Oversampling for multiclass sentiment analysis on Amazon Review Dataset.* 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), Morioka, Japan, 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923260

Rohan Badlan. 2019. *Disambiguating Sentiment: An Ensemble of Humour, Sarcasm, and Hate Speech Features for Sentiment Classification.* Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 337–345

Anna Rogers, Olga Kovaleva, Anna Rumshisky. 2020. *A Primer in BERTology: What we know about how BERT works. ArXiv* abs/2002.12327