Sentiment Analysis of Amazon Product User Review using Natural Language Processing (NLP)

By Nirajan Bhattarai, CS6699

Abstract

Natural language processing (NLP) is a special domain comprising computer science, artificial intelligence and linguistics, where computer algorithms try to understand and manipulate natural language expressions—knowledge, emotion or exchange of information between people or people and environment. Sentiment analysis in Natural language processing is simply a way to process the input to label underlying emotions attached to that expression. In this study, we tried to derive a model, based upon the amazon review dataset of mobile electronics, that can successfully classify a given product review to a particular sentiment inclination. We implemented supervised learning with GRU and LST, unsupervised learning with VADER, and transfer learning with Hugging Face BERT model. The supervised learning faced high predictions in false positive and false negative region, and transfer learning with BERT suffered heavily with overfitting, and accuracy was just 67%. Nevertheless, VADER sentiment classification showed strong agreement with label rating of the review. So, VADER is considered the best overall model, which is unsupervised as well, to classify a product review polarity associated with mobile electronics.

Keywords **—model, sentiment, supervised/unsupervised, review**

I. Introduction

Online trading/shopping platforms are now a dominant marketplace, which offers convenience, diversity, and time saving to the consumer. Product, service or brand review dominates the buying behavior of the consumer. The value a Natural Language Processing (NLP) can bring to this scenario is automatic rapid processing of the huge amount of reviews about a product and label a user inclination to that product, services or a brand (Dey S et al, 2020).

Obtaining the hidden sentiment on any form of expression/communication has a wide range of tactical and strategic significance in interpersonal relationships. With the advent of the internet, there is a surge of large volumes of user data, and a significant chunk of this is a user review on the purchases they made, services they subscribe to, shows they watch and so on (Gosh A, 2018). In business domain, it provides insight about subjective perception or sentiment of the consumer toward the product, thus allowing business to be responsive to consumer feedback, act on negative feedback, and monitor brand reputation in a real time (Nadkarni et al, 2011). The problem of processing reviews or opinions is their unstructured or semi-structured nature, thus very difficult for human intelligence, which is heavily dependent on structured information for a particular inference, to process.

Machine learning can efficiently process these tasks in a short period of time, and with relatively less bias. However, there are challenges. Being unstructured or semi-structured, data pre-processing requires removal of punctuation, lemmatization, and tokenizing the vocabulary using filters. Inappropriate preprocessing may lead to faulty interpretation (Nadkarni et al, 2011). The main objective of this project is to apply natural language processing (supervised and unsupervised learning) to the amazon review dataset to generate a model that can label any future review to a particular emotional inclination. Supervised, unsupervised and transfer learning techniques were applied to modeling of a amazon review dataset.

II. Background

Natural language processing (NLP) is a special domain comprising computer science, artificial intelligence and linguistics, where computer algorithms try to understand and manipulate natural language expressions—knowledge, emotion or exchange of information between people or

people and environment. NLP is a way to extract grammatical structures and semantics from the input, and apply that to useful tasks like language generation, emotion identification, translation, so on and so forth (Reshwala A, Mishra D and Pawar P, 2013; Chowdhary K, 2013).

Sentiment analysis in Natural language processing is simply a way to process the input to label a underlying emotions attached to that expression (Rajput A, 2020). Sentiment analysis sometimes is intriguing because of the different conditioning of cultures and society on humans, which forces them to perceive the same reality differently. Also, sometimes expressions can be abstract, and there can be up and down movement of sentiment within a single subject, which is very hard to discern and label a specific emotion to it.

User sentiment towards a product, service or brands plays a huge role in its success or failure. A contextual text mining which is extractive of subjective social sentiment, in particular, of a brand, products or services from a provided source material like online conversation, reviews, rating etc. may be perceived as a sentiment analysis in the industry(Gupta S, 2018).It is of no doubt that a product, services or brand attributes, which dictates particular dominant sentiment among consumers, heavily influence its success, especially in a online space, where actual physical perception and realization is limited. A consumer sentiment may be a direct measure of their satisfaction of a product, services or brand (brand24.com, 2018).

III. Method

A. Dataset

The amazon review dataset for sentiment modeling and analysis was obtained from https://www.tensorflow.org/datasets/catalog/amazon_us_reviews loaded as TFDS dataset with 'train' split.

B. Preprocessing

Preprocessing included extracting text and sentiment from the entire dataset, encoding sentiment, cleaning text review, and then tokenization and vectorization. Natural language toolkit (NLTK), tensorflow text vectorization, and BertTokenizerFast were used for tokenizing and vectorizing the text based upon the model implied. The final step was preparation of the data suitable to feed into the model.

C. Models

Dataset was trained into three different model architectures, which included traditional Gated Recurrent Neural Network (GRU) with embeddings and text vectorization layer, VADER unsupervised text classification model, and transfer learning with Hugging Face distilbert-based-uncased(https://huggingface.co/distilbert-base-uncased).

D. Model Evaluation

Model evaluation was carried out using accuracy matrices, and confusion matrix. Training and validation dataset were used for training and its while test split will be used to evaluate the model.

IV. Result and Discussion

The dataset was tested using supervised, unsupervised, and transfer learning models.

A. Data Exploration

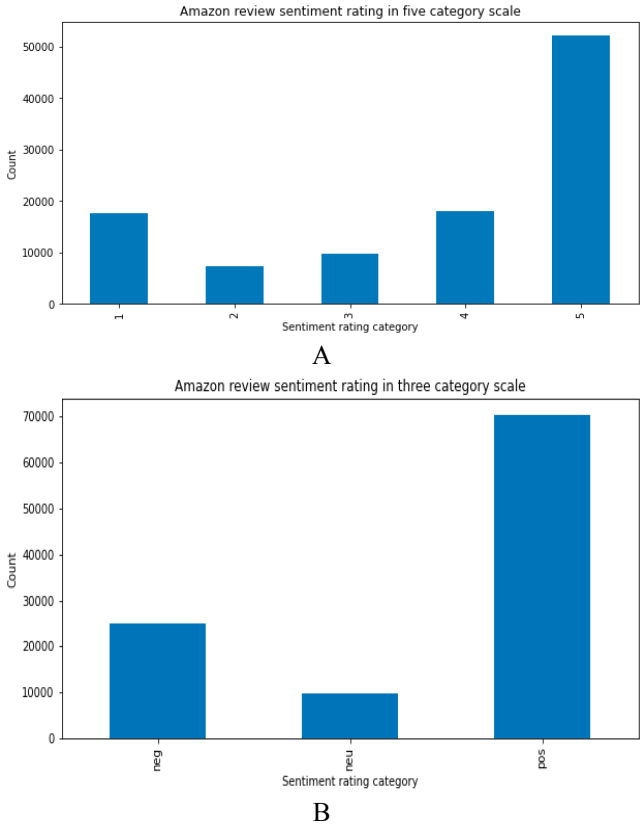
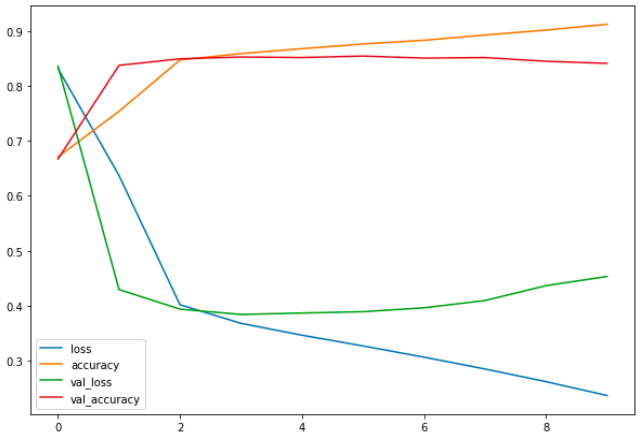


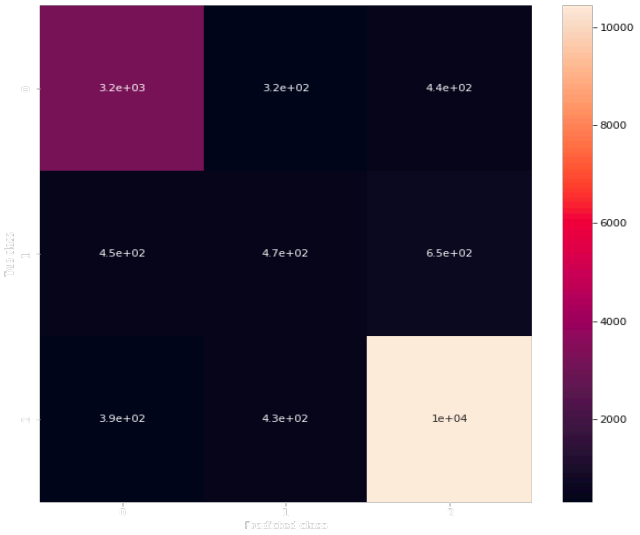
Figure 1. Figure showing distribution review sentiment rating in 3-point(A) and 5-point scale(B)

The rating was unevenly distributed, more weighted toward positive at the extreme end. In three-point scale, we assume numeric review rating 1-2 to be positive, 3 being neutral and 4-5 being positive. The 5-point scale is presented as it is in the provided dataset. Results showed the dataset is unbalanced and skewed toward the positive rating. The possible solutions to deal with unbalanced dataset are under sampling, oversampling or synthetic sampling but it will either significantly eliminate important information for model training or distort it. Thus, the model may be biased toward positive rating, which, we think, is a limitation to this study. We feed the model with a dataset as such.

B. Supervised learning models (GRU and LSTM) models



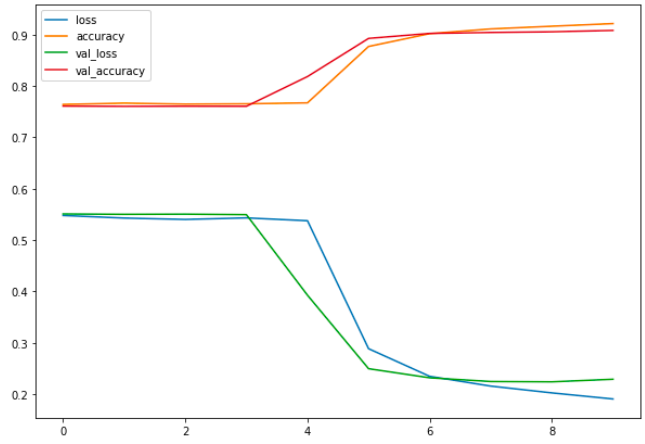
A



B

Figure 2. History plot of train and validation accuracies(A) and confusion matrix(B) for GRU model

The first model we implemented was Gated Recurrent Network (GRU) with text vectorization layer, and word embeddings as a supervised learning model. The result showed the model is fairly fit as shown in the history plot, and test accuracy is 90%. In confusion matrix, significant observations were found to be false positives and false negatives.



A

B

Figure 3. History plot of train and validation accuracies(A) and confusion matrix(B) for LSTM model

Table1: Table showing combined accuracies for GRU and LSTM models

Accuracy	**(GRU)**	**(LSTM)**
Train Accuracy	86%	86%
Validation Accuracy	84%	84%
Test accuracy	90%	90%

The LSTM model mimicked the output of GRU in all accuracy evaluation matrices.

C. Unsupervised VADER model

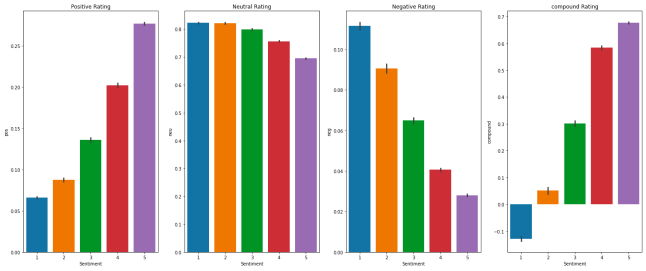


Figure 4. Figure showing the distribution of different VADER score against review rating label

VADER, Valence Aware Dictionary for Sentiment Reasoning, is an unsupervised sentiment analysis model which classifies given text based upon polarity and intensity of its emotional bearing. It maps lexical features to emotion intensities, known as sentiment score, and then sums up intensities of all words to derive particular inferences. It has positive, neutral, negative and compound sentiment scores (Beri, 2020). Compound score is a normalized sum score of positive, negative and neutral, -1 is ultimate negative and +1 is ultimate positive rating of the text (Analytics India Magazine, 2020). The result has shown strong correlation between emotional intensity scores assigned by the VADER model as compared to their respective sentiment labels. For instance, intensity rating 5 has also got the highest average positive score from VADER, and review, which was labeled as rating 1, received highest average negative sentiment score from VADER. Compound scores also showed agreement with sentiment rating labels of the reviews. The average compound score for rating 1 is negative, and the trend is that the average compound score increases as the rating score increases.

D. Transfer Learning with Hugging Face BERT model

Developed in google, BERT, bidirectional Encoder Representation from Transformers, is trained in 2500 words for English Wikipedia and 800 M BooksCorpus words, has a base model architecture of 12 layer, 768 hidden, 12 head, and 110M parameters (Yalçın O.G, 2021). The result showed the Hugginge Face Bert Model suffered heavily with overfitting problem with train and validation accuracy of just 66-67%. This is a transfer learning and we expected the model would be far better than this. The reason for this poor performance might be due to improper hyperparameter tuning or inappropriate preprocessing of the dataset before it is fed into the model.

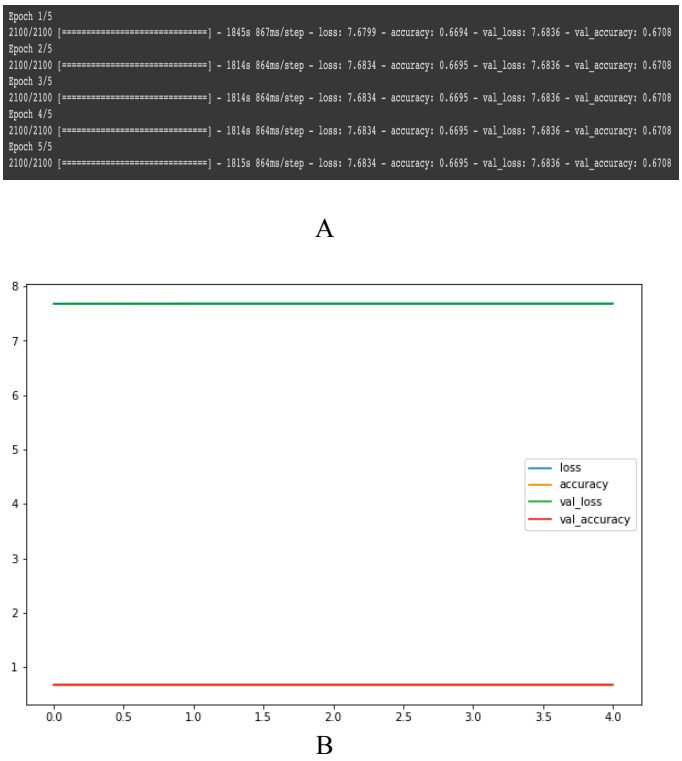


Figure 5. Figure showing the training history(A) and history plot for Hugging face Bert Model(B)

V. Conclusion

We applied supervised learning with GRU and LSTM, and both mimicked each other. The respective train, validation accuracy was 86,84, and 90%, the models were fairly fit, but the predictions were mostly in the region of false positives and false negatives. In unsupervised learning with VADER, the sentiment polarity analysis of the reviews was consistent with their respective rating label. Review rating of 5 has highest average positive score, while review rating 1 has lowest average negative score, and compound score was negative for rating of 1 and has highest positive score for review with rating 5. The transfer learning with hugging face BERT model did not perform well as expected. The model was highly overfit with accuracy of 66-67%. In overall, we propose VADER classification of review text is best overall model

References

Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J. and Dey, M., 2020, February. A comparative study of support vector machine and Naive Bayes classifier for sentiment analysis on Amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)* (pp. 217-220). IEEE.

Rajput, A., 2020. Natural language processing, sentiment analysis, and clinical analytics. In *Innovation in Health Informatics* (pp. 79-97). Academic Press.

Reshamwala, A., Mishra, D. and Pawar, P., 2013. Review on natural language processing. *IRACST Engineering Science and Technology: An International Journal (ESTIJ)*, *3*(1), pp.113-116.
Chowdhary, K., 2020. Natural language processing. *Fundamentals of artificial intelligence*, pp.603-649.

Ghosh, A., 2022, March. Sentiment Analysis of IMDb Movie Reviews: A comparative study on Performance of Hyperparameter-tuned Classification Algorithms. In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 289-294). IEEE.

Nadkarni, P.M., Ohno-Machado, L. and Chapman, W.W., 2011. Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), pp.544-551.

Beri, A. (2020). *SENTIMENTAL ANALYSIS USING VADER*. [online] Medium. Available at: https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664.

Analytics India Magazine. (2020). *Sentiment Analysis Made Easy Using VADER*. [online] Available at: https://analyticsindiamag.com/sentiment-analysis-made-easy-using-vader/#:~:text=The%20compound%20score%20is%20the