ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 2, pp. 101-110
DOI:10.33168/JSMS.2023.0207

Comparing Machine Learning and Deep Learning Based Approaches to Detect Customer Sentiment from Product Reviews

Sin Li Lim, Lee Kien Foo, Sook-Ling Chua
Multimedia University, Cyberjaya, 63100, Malaysia
lkfoo@mmu.edu.my (Corresponding Author)

Abstract. Product review is way for customers to express their sentiments towards a product. Sentiment analysis can be performed to gain insights from product reviews and help to improve the sale-ability of a product. This research aims to perform sentiment analysis on reviews of electronic products. In our study, we compared two methods: one is based on deep learning and another is based on machine learning. The public data of product reviews obtained from Amazon and BestBuy were used in this study. We compared a deep learning method and a machine learning method by first investigated the impact of training data selected using different sampling techniques. Then we examined the effect of hyper-parameter tuning on these learning algorithms. The resulted best models were then compared to the baseline model for sentiment analysis. The experimental results showed that our models performed better than the baseline model in terms of accuracy and F1. The experimental outcome suggests that customer's sentiment towards a product can be captured by applying deep learning and machine learning on product reviews.

Keywords: Sentiment analysis, convolutional neural network, sampling techniques, support vector machine.

1. Introduction

In the past years, there are increase numbers of users purchase items online via e-commerce websites such as Amazon, BestBuy and Flipkart. Sultana et al. (2019) reported that it is common for users to review their purchases by sharing their views and feelings. These reviews are important not only to the companies but also to the potential customers (Kasturia et al. 2020). By tracking the customer sentiments, the company can improve its product and customer experiences. However, Guerreiro and Rita (2020) showed that with the increasing number of reviews for each product, it is indeed a tedious and time-consuming process to manually filter out the reviews in order to understand the customer sentiments. Recent studies have applied machine learning and deep learning to automatically learn and polarize the product reviews (Haque et al. 2018, Liu et al. 2020, Sharma and Jain 2020).

One of the challenges in analysing product reviews is that these reviews often consist of mixed emotions and ratings. Sometimes, reviews from the customers were very bad but a high rating were given (e.g., 5 stars); other times reviews on the products were good but ratings were very low. It is difficult to determine the 'true' feelings of the customers giving such reviews and ratings. Sentiment analysis is one of the commonly used techniques to determine the sentiment in a post. Machine learning and deep learning have been applied on sentiment analysis but majority of the studies focus on only positive and negative sentiments. Since customer may not have an extreme positive or negative feeling towards a product, we have included the neutral sentiment in our study and perform multiclass classification. We have investigated the impact of different sampling techniques in the selection of training data and examined the effect of hyper-parameter tuning on the learning algorithms.

The remainder of this paper is organized as follows. Section 2 reviews the related work on sentiment analysis. Section 3 describes the datasets, learning algorithms and sampling techniques applied in this study. Section 4 presents our experimental setup. Section 5 discusses the experimental results and Section 6 concludes our findings.

2. Related works

Sentiment analysis, also known as polarity detection, is the process to detect positive or negative sentiment in text (Sun et al. 2019). Sentiment analysis is commonly used to understand how customers feel toward a product. Several approaches have been proposed in the literature for sentiment analysis.

In the work of Guia et al. (2019), they compared different machine learning methods for sentiment analysis. Their study found that support vector machine (SVM) achieved a better performance than naïve Bayes, random forest and decision trees. A similar study was conducted by Dey et al. (2020) where they compared SVM with naïve Bayes classifier. Their work analysed the sentiments of product reviews from Amazon and findings showed that SVM has better accuracy. Imamah et al. (2020) applied SVM for sentiment analysis on tourist reviews. The reviews were in Indonesian language extracted from Tripadvisor. Traditional sentiment analysis methods, however, do not consider the ambiguous meaning of a word in a sentence. To address this problem, Song et al. (2020) referred to the word ambiguity as polysemy and proposed a text representation model named Word2PLTS for short text sentiment analysis by introducing probabilistic linguistic terms sets. They also created a sentiment analysis and polarity classification framework named SAPCP by using SVM.

There are works that applied deep learning for sentiment analysis. The works of Socher et al. (2011 and 2013) are among the earlier works that used deep learning. In these studies, they proposed a recursive neural tensor network model to analyse the compositional effects of sentiment in language. Ouyang et al. (2015) used the convolutional neural network (CNN) to classify sentiments on sentences from movie reviews. Findings from Wang et al. (2016) showed that sentiment polarity of a sentence is related to the aspect. They

proposed an attention-based long short-term network for aspect-level sentiment classification. Their method learns aspect embedding, which is used to determine the attention weights.

Yang et. al. (2020) proposed a sentiment analysis model based on CNN and recurrent neural network. They first calculate the weights of words using a sentiment dictionary. CNN and gated recurrent unit are then applied for feature extraction and sentiment classification. Their work was evaluated on product reviews from a Chinese e-commerce website. Although their model enhanced the sentiment features of the input text, it may not be suitable when high sentiment refinement is required. Poomka et. al. (2021) compared the classification performance of machine leaning and deep learning algorithms on product reviews for positive and negative sentiments. They studied the effect of data preprocessing and found that the preprocessing step improved the performance of machine learning algorithms, but not deep learning.

Study in sentiment analysis tends to focus on the positive and negative sentiments. The neutral class is assumed to be equivalent to no opinion in the sentiment level (Liu 2012). Recently, researchers have recognized the importance to include the neutral sentiment in their studies (Al-Rubaiee et al. 2016, Kumar et al. 2019, Roccabruna et al. 2022, Nurkholis et al. 2022). However, the learning algorithms showed mixed performance with the inclusion of neutral class. In this study, we investigated the impact of training data selected by using different sampling techniques on the performance of machine learning and deep learning algorithms. We also investigated the effect of hyper-parameter tuning on the performance of learning algorithms when neutral sentiment is included.

3. Methodology

The description of our research methodology is provided in this section. Section 3.1 is the description of the datasets used in our study. The machine learning and deep learning algorithms is described in section 3.2. The sampling techniques investigated in this study is explained in section 3.3.

3.1 Datasets and data pre-processing

Two datasets were used in this study. The first dataset was obtained from Amazon and Best Buy Electronics. The second dataset was obtained from Ni et al. (2019). Both datasets contain reviews on electronic products. These datasets were combined resulting a total of 12126 records of product reviews.

From the dataset, each review is manually labelled and the final count is presented in Table 1. The reviews were pre-processed before model training. The steps include: (1) transforming all words to lowercase, (2) removal of contractions, special characters, punctuations, numbers, stop-words and (3) tokenization. Further processing depends on the classification algorithms (Section 3.2).

Table 1: Total number of records for each class

Sentiment	No. of Records
Positive	6172
Neutral	2576
Negative	3378

3.2 Classification algorithms

We have evaluated one machine learning algorithm and one deep learning algorithm in this study. SVM is the machine learning algorithm chosen in our study since existing literatures reported that SVM performed well for sentiment analysis (Guia et al. 2019, Dey et al. 2020). The dataset was further process into vectorised format using term-frequency inverse document frequency for SVM. The radial basis function was selected in our study as the SVM kernel for the handling of nonlinear multiclass classification. There are two hyper-parameters in SVM, the C parameter that control the margin of the decision boundary and

the gamma parameter that control the influence of the support vectors on the position of the hyper-lines. Hyper-parameter tuning was performed on these two parameters by using grid search method.

Deep learning is the application of artificial neural networks to learn tasks using multiple-layered networks (Zhang et al. 2018). We have selected the CNN in this study since CNN is proven to be more computationally efficient compared to other deep learning algorithms (Ouyang et al. 2015). The data is converted to word embedding format using word2vec and padded with the sequence when applied CNN. The architecture of our CNN model is shown in Figure 1. The CNN was trained with 100 epochs and a batch size of 64. Hyper-parameters that are common in CNN includes the number of hidden layers, dropout, initialization of network weights, activation functions, learning rate, number of epochs, and batch size. we used the Keras tuner library (Malley et al. 2019) to perform hyper-parameter tuning in CNN.

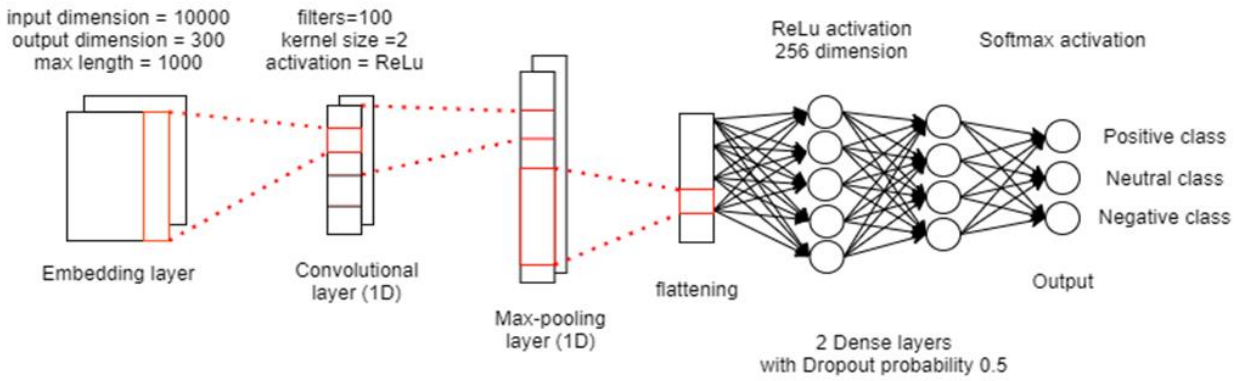

Fig.1: Architecture of our CNN model

3.3 Sampling techniques

Two sampling techniques were investigated in this study. The first sampling technique is the balanced sampling where the training data consists of equal number of instances from each sentiment class. The learning algorithm is expected to have the same amount of training instances to learn for each class with the balanced sampling technique. In our experiments, we only balanced the training data and the testing data will consist all the remaining instances. The second sampling technique is the stratified sampling where the training data is selected according to the overall data ratio of each class, i.e., if 40% of instances in the original dataset is class A then the training data will consist 40% class A instances. The training data selected with stratified sampling preserves the proportions of data in each class as in the original dataset.

4. Experimental setup

We conducted three experiments. In each experiment, two models were trained for sentiment analysis: (1) CNN and (2) SVM. The performance of the models was evaluated in terms of classification accuracy and F1. We applied 5-fold cross-validation for each experiment. Figure 2 shows the overall process of our experiments.

4.1 Experiment 1: sampling techniques

The first experiment is to investigate the impact of different sampling techniques on the classification performance. For balanced sampling, 1000 instances were selected from each sentiment class to form the training data. For stratified sampling, we sampled the instances according to the proportion of class size. The training data consists of 3000 instances with either sampling technique. All the remaining instances were used as test data. The partition of training-test data is detailed in Table 2. In this experiment, two CNN

models were trained -- one on training data selected with balanced sampling and the other on training data selected with stratified sampling. We did the same with SVM.

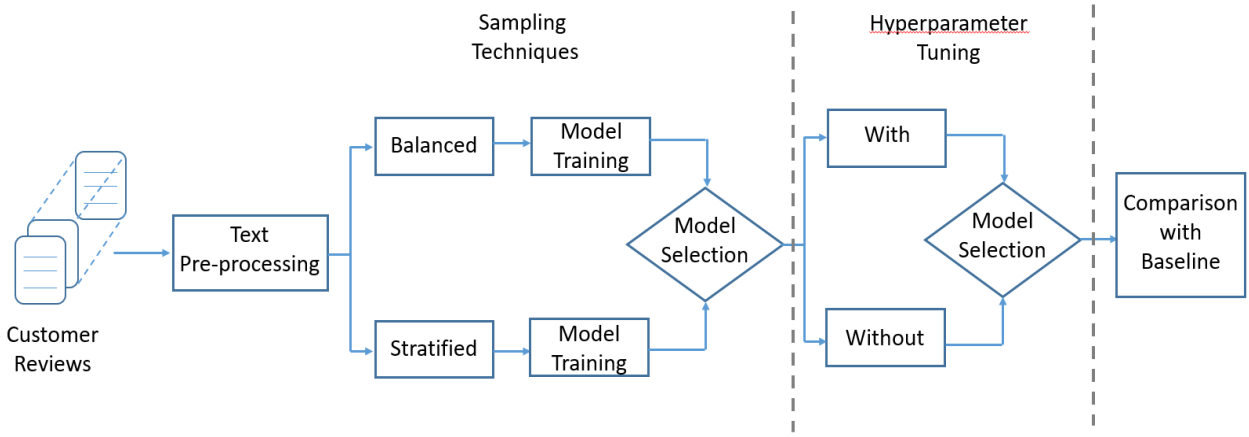

Fig.2: Illustration on the process of our experiments

Table 2: Chi-square test result

Sentiment	Balanced Sampling		Stratified Sampling	
	Training	Testing	Training	Testing
Positive	1000	5172	1527	4645
Neutral	1000	1576	637	1939
Negative	1000	2378	836	2542
Total	3000	9126	3000	9126

4.2 Experiment 2: hyper-parameter tuning

Hyper-parameters are said to play a crucial role in the performance of both CNN and SVM (Andonie 2019). Our second experiment is to investigate if hyper-parameter tuning improves the performance of CNN and SVM for sentiment analysis. In this experiment, we have performed 'randomsearch' and 'hyperband' hyper-parameter tuning on CNN. The 'randomsearch' method selects the combination of hyper-parameter values randomly, and then performs full training and evaluates each of the combination. The 'hyperband' method, on the other hand, randomly samples all the combinations of hyper-parameter and then selects the best candidates set based on the initial training of the model for a few epochs (Li and Jamieson 2018). For SVM, the optimal hyper-parameters can be found by creating a grid of hyper-parameters and try all the possible combinations. We applied the 'gridsearch' function in the Python Scikit-learn library for hyper-parameter tuning on SVM.

4.3 Experiment 3: comparison with baseline method

Our third experiment is to compare the performance of CNN and SVM with the baseline Valence Aware Dictionary and sEntiment Reasoner (VADER). VADER is a rule-based sentiment analysis tool proposed by Hutto and Gilbert (2014). In this experiment, the model that performed the best from our previous experiments (one for CNN and one for SVM) will be selected and compared to VADER.

5. Results and discussion

5.1 Experiment 1: sampling techniques

The performance of CNN with training data selected using the two sampling techniques are presented in Figure 3. The grouped bar plots showed the comparison of the average accuracy and average F1 of the 5-

fold cross-validation for all the three classes. As shown in Figure 3, CNN performs better in both accuracy and F1 across all the classes when trained with data sampled using balanced sampling technique.

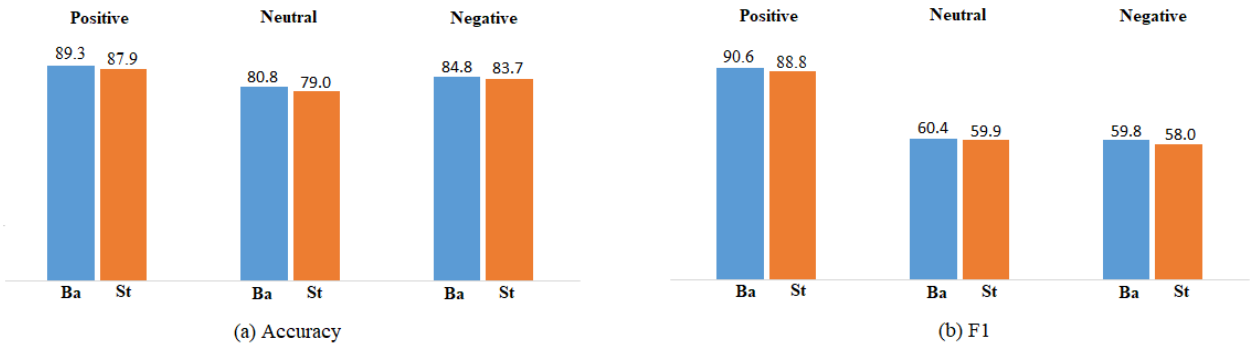

Fig. 3: Classification performance of CNN with balanced (Ba) and stratified (St) sampling.

The classification performance of SVM with different sampling techniques are shown in Figure 4. Unlike CNN, there is no clear winner for SVM with balanced or stratified sampling. In terms of accuracy, SVM trained with data selected using stratified sampling performed better in positive class but SVM trained with data selected using balanced sampling perform marginally better in the neutral and negative classes. In terms of F1, SVM trained with data selected using stratified sampling performed better in both positive and neutral classes, but not as good in the negative class. The difference is marginal except for the neutral class.

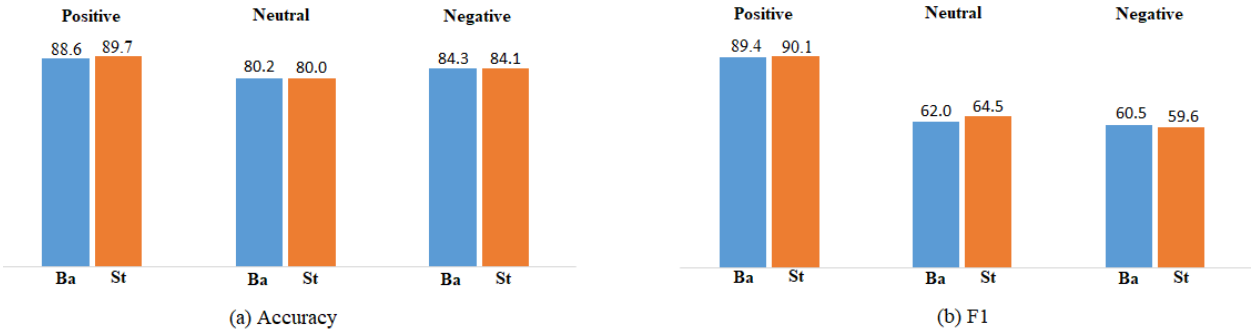

Fig. 4: Classification performance of SVM with balanced (Ba) and stratified (St) sampling

5.2 Experiment 2: hyper-parameter tuning

CNN trained with data selected using balanced sampling performed better in experiment 1 and therefore this model is used to evaluate the effect of hyper-parameter tuning. The 'randomsearch' and 'hyperband' methods were used to perform hyper-parameter tuning for CNN. The results in Figure 5 showed that hyper-parameter tuning did not have an effect on the classification performance of CNN. CNN without hyper-parameter tuning performed better across all the classes in terms of accuracy and F1, as shown in Figure 5.

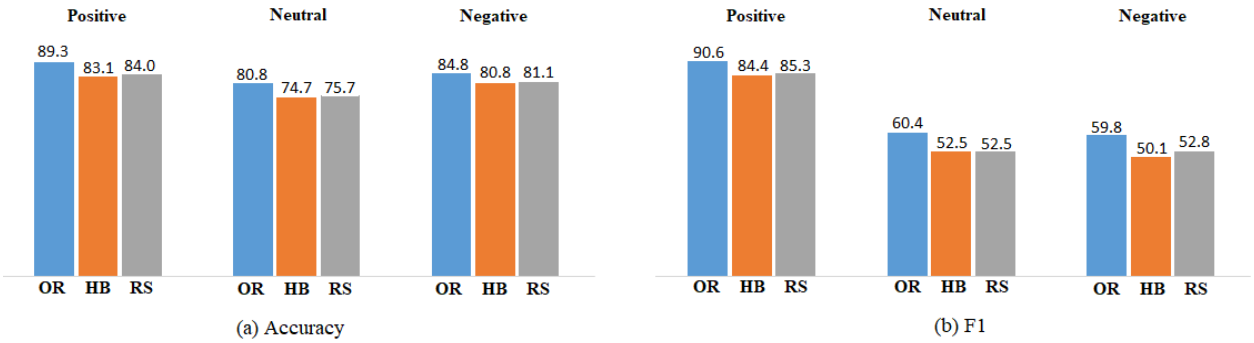

Fig.5: Classification performance of CNN trained on data selected using balanced sampling: (1) without hyper-parameter tuning (OR), (2) with 'hyperband' hyper-parameter tuning (HB) and (3) with 'randomsearch' hyper-parameter tuning (RS).

Although there is no clear winner for SVM in experiment 1, we have chosen the SVM trained with data selected using stratified sampling since it performed slightly better in general. We compared the performance of this model with the hyper-parameter tuning SVM model. The hyper-parameter tuning is carried out by applying a grid-search algorithm on the hyper-parameter space. Figure 6 shows that SVM with hyper-parameter tuning performed slightly better in both accuracy and F1 for all the classes.

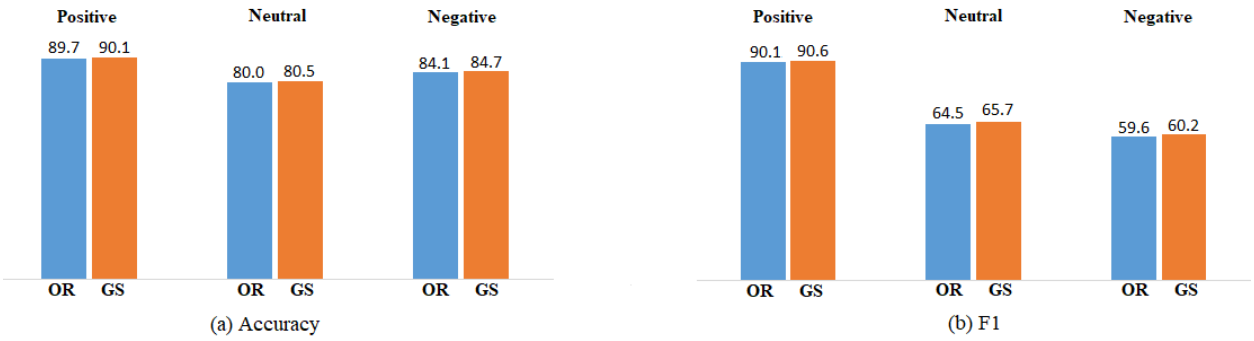

Fig. 6: Classification performance of SVM trained on data selected using stratified sampling: (1) without hyper-parameter tuning (OR) and (2) with 'grid search' hyper-parameter tuning (GS).

5.3 Experiment 3: comparison with baseline method

This experiment compares the best model of CNN (balanced sampling without hyper-parameter tuning) and SVM (stratified sampling with hyper-parameter tuning) to the baseline method, VADER. The results are presented in Table 3. The best result for each class is in bold.

Referring to Table 3, the best model of CNN and SVM performed better than VADER in both accuracy and F1 across all the classes. The result showed that both the CNN and SVM outperformed VADER for sentiment analysis. When comparing CNN to SVM, SVM performed better in accuracy for positive class, but not as good for neutral and negative classes. In term of F1, CNN and SVM performed equally well for positive class (90.6%). For neutral and negative classes, SVM performed better than CNN.

Table 3: Comparison between CNN, SVM and VADER in terms of accuracy and F1

Method	Accuracy			F1		
	Positive	Neutral	Negative	Positive	Neutral	Negative
CNN	89.3	**80.8**	**84.8**	**90.6**	60.4	59.8
SVM	**90.1**	80.5	84.7	**90.6**	**65.7**	**60.2**
VADER	65.4	72.2	82.7	74.3	10.8	49.6

6. Conclusions

Sentiment analysis on product reviews can provide important insights to understand the customer viewpoints and their sentiments toward a product. In this research, we have conducted experiments to analyse the sentiments on electrical product reviews with CNN and SVM. We examined the impact of training data sampled using balanced and stratified sampling techniques. The results showed that CNN performed better when the training data is sampled with balanced sampling technique, while SVM has mixed performance. We have also investigated the effect of hyper-parameter tuning on CNN and SVM for sentiment analysis. The results showed no significant improvement with hyper-parameter tuning in CNN and marginal improvement in SVM. The best model of CNN and SVM were compared to the baseline method (VADER) for sentiment analysis. Both CNN and SVM performed better than the baseline method.

In this study, we have applied machine learning and deep learning in sentiment analysis with the focus on the sampling techniques and hyper-parameter tuning. From the experimental results, the overall performance of the multiclass sentiment analysis models built with either machine or deep learning algorithms are reasonably well and can be used to predict the customer sentiment from product review. We found that the sampling technique influence the performance of the deep learning algorithm but has no significant effect on the machine learning algorithm. On the other hand, hyper-parameter tuning improve the performance of the machine learning algorithm but not the deep learning algorithm. However, we have only evaluated one machine learning and one deep learning algorithm in this study, for future work we plan to extend our study to other learning algorithms.

***Funding Statement*:** The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Sultana N., Kumar P., Patra M. R., Chandra S. & S. Alam (2019). Sentiment Analysis for Product Review. *ICTACT Journal on Soft Computing*, 9, 1913–1919.

Kasturia V., Sharma S., & Sharma S., (2020). Automatic Product Saleability Prediction Using Sentiment Analysis On User Reviews. *International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020.

Guerreiro J. & Rita P. (2020). How to Predict Explicit Recommendations in Online Reviews Using Text Mining and Sentiment Analysis. *Journal of Hospitality and Tourism Management*, 43, 269–272.

Haque T. U., Saber N. N. & Shah F. M.,(2018). Sentiment Analysis on Large Scale Amazon Product Review. *IEEE International Conference on Innovative Research and Development (ICIRD)*, 2018.

Sharma S. & Jain A. (2020). Role of Sentiment Analysis in Social Media Security and Analytics. *WIREs Data Mining and Knowledge Discovery*, 10, e1366.

Liu Y., Lu J., Yang J. & Mao F. (2020). Sentiment Analysis for E-Commerce Product Reviews by Deep Learning Model of BERT-Bigru-Softmax. *Mathematical Biosciences and Engineering*, 17, 7819–7837.

Sun Q., Niu J., Yao Z. & Yan H. (2019). Exploring eWOM in Online Customer Reviews: Sentiment Analysis at a Fine-Grained Level. *Engineering Applications of Artificial Intelligence*, 81, 68–78.

Guia M., Silva R. & Bernardino J., (2019). Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management (IC3K)*, 2019.

Dey S., Wasif S., Tonmoy D., Sultana S., Sarkar J. & Dey M.,(2020). Comparative Study of Support Vector Machine and Naïve Bayes Classifier for Sentiment Analysis on Amazon Product Reviews. *International Conference on Contemporary Computing and Applications (IC3A)*, 2020.

Imamah, Husni, Rachman E., Suzanti I. & Mufarroha F., (2020). Text Mining and Support Vector Machine for Sentiment Analysis of Tourist Reviews in Bangkalan Regency. *Journal of Physics: Conference Series (JPCS)*, 2020.

Socher R., Lin C. C., Ng A. Y. & Manning C. D.,(2011). Parsing Natural Scenes and Natural Language with Recursive Neural Networks. *International Conference on Machine Learning (ICML)*, 2011.

Socher R., Perelygin A., Wu J., Chuang J., Manning C., Ng A. & Potts C., (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Ouyang X., Zhou P., Li C. H. & Liu L.,(2015). Sentiment Analysis Using Convolutional Neural Network. *IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM)*, 2015.

Wang Y., Huang M., Zhao L. & Zhu X.,(2016). Attention-Based LSTM for Aspect-Level Sentiment Classification. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang L., Li Y., Wang J. & Sherratt R. S. (2020). Sentiment Analysis for E-Commerce Product Reviews in Chinese based on Sentiment Lexicon and Deep Learning. *IEEE Access*, 8, 23522–23530.

Poomka P., Kerdprasop N. & Kerdprasop K. (2021). Machine Learning Versus Deep Learning Performances on The Sentiment Analysis of Product Reviews. *International Journal of Machine Learning and Computing*, 11, 103–109.

B.Liu (2012). Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies*, 5, 1–167.

H. Al-Rubaiee, R. Qiu & D. Li,(2016). The Importance of Neutral Class in Sentiment Analysis of Arabic Tweets. *International Journal of Computer Science & Information Technology (IJCSIT)*, 2016.

S. Kumar, M. Yadava & P. P. Roy (2019). Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *Information Fusion*, 52, 41–52.

G. Roccabruna, S. Azzolin & G. Riccardi, (2022). Multi-source Multi-domain Sentiment Analysis with BERT-based Models. *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, 2022.

A. Nurkholis, D. Alita & A. Munandar (2022). Comparison of Kernel Support Vector Machine Multi-Class in PPKM Sentiment Analysis on Twitter. *Jurnal Rekayasa Sistem dan Teknologi Informasi*, 6(2).

Datafiniti, Amazon and Best Buy Electronics, Available: https://data.world/datafiniti/amazon-and-best-buy-electronics.

Ni J., Li J. & Mcauley J.,(2019). Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Zhang L., Wang S. & Liu B. (2018). Deep Learning for Sentiment Analysis: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), 1–25.

Malley T., Bursztein E., Long J. & Chollet F., Keras Tuner. https://github.com/keras-team/keras-tuner 2019.

Andonie R. (2019). Hyper-parameter Optimization in Learning Systems. *Journal of Membrane Computing*, 1, 279–291.

Li L. & Jamieson K. (2018). Hyperband: A Novel Bandit-Based Approach to Hyper-parameter Optimization. *Journal of Machine Learning Research*, 18, 1–52.

Hutto C. & Gilbert E., (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of social media Text. *International AAAI Conference on Weblogs and social media (ICWSM)*, 2014.