Model Comparison in Sentiment Analysis: A Case Study of Amazon Product Reviews

Honglie Zhang*

University of Alberta, Edmonton, Alberta, T6G 2C6, Canada

*Corresponding author email: honglie@ualberta.ca

Abstract. Sentiment analysis is essential in NLP, especially in businesses because it can improve customer services. This paper focuses on a particular case of sentiment analysis, a case study of Amazon reviews of books on kindle. Firstly, this paper applies several non-deep-learning algorithms including Logistic Regression, Naïve Bayes, Support Vector Machine, Convolutional Neural Network, and Recurrent Neural Network, and compares their accuracies. Especially, for deep learning methods, this paper studies the slope of accuracies concerning the number of hidden layers. Secondly, as a multi-class text classification problem, the product review data set has five labels ranging from one star to five stars, a new method called Hybrid Sequential Binary Classification (HSBC) is introduced in this paper, which improves the behavior of classical binary classifiers on a multi-class text classification problem. Moreover, a comparison of HSBC and multi-class classification models is presented.

Keywords: Natural Language Processing, Multi-class Classification, Model Comparison, Hybrid Sequential Binary Classification (HSBC).

1. Introduction

Natural Language Processing (NLP) makes efforts to help computers access linguistic intelligence. As a special topic of NLP, sentiment analysis deals with the emotional implication of texts. A comprehensive survey of sentiment analysis is given in the exploration of Walaa M et al. [1]. Classically, sentiment analysis is used to categorize comments on social media into the general frame: "Negative" or "Positive", and sometimes "Neutron" is included as well. Common applications consist of the study of Vishal A. Kharde and S.S. Sonawane on Twitter data analysis [2], research of B. Bhutani et al. on fake news filtration [3], etc. However, sometimes classification of this type is not enough, for instance, the multi-class classification problem is studied in the work of Alsafy et al. [4] where the cardinality of label class is usually greater than two. Product review is a classification problem of this type, and there are already several attempts, for example, Fang, X. and Zhan, J. have applied Sentiment polarity categorization for Amazon product reviews [5] and Abinash T et al. have solved multi-class classification on reviews using machine learning [6].

This paper focuses on a case study of Amazon kindle product reviews where product reviews consist of both comments and star evaluations ranging from one to five. Our task is to classify each comment into its corresponding star evaluation. As a multi-class classification problem, this study focus on the performance of several machine learning classifiers concerning two categorizations I and II, where I={"Negative", "Neutron", "Positive"} and II={1,2,3,4,5}.

Besides the classical algorithms containing Logistic Regression, Naïve Bayes and Support Vector Machine (SVM), the popular deep learning techniques are also taken into account. Several comparisons of models are presented:

(1) The comparison of the performance of models concerning two categorizations I and II.

(2) The comparison of the slope of the accuracy of deep learning models concerning the number of hidden layers.

Especially, as multi-class classification problems can be sliced into several sub-classification problems, this paper introduces a new method called Hybrid-Sequential-Binary-Classification (HSBC) where different binary classifiers are applied in sub-binary classification. It is a class of several combinatorial models and a comparison of accuracies of all models in this paper is included.

2. Methodology

In the case study of Amazon kindle book reviews, the data set is taken from Kaggle (https://www.kaggle.com/datasets/meetnagadia/amazon-kindle-book-review-for-sentiment-analysis), which consists of 12000 book reviews including detailed comments and star evaluation ranging from 1 star to 5 stars. Quantitative and Proportional distributions of star evaluations are presented in Figure 1 and Figure 2 respectively as follows:

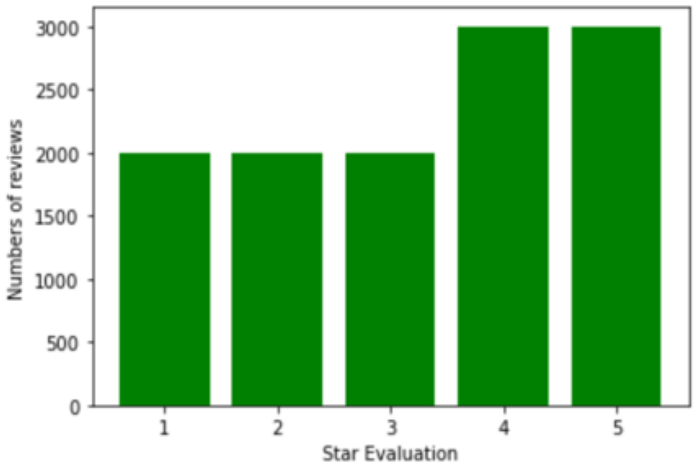


Figure 1. Quantitative Distribution

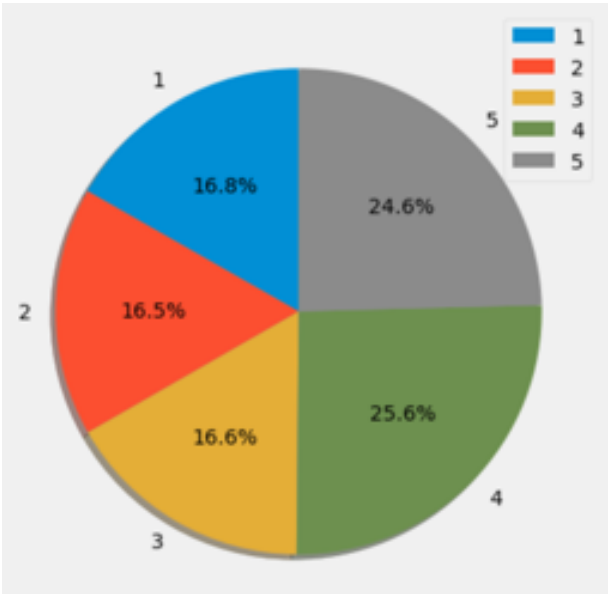


Figure 2. Proportional Distribution

For this particular multi-class classification problem, two categorizations are applied to studying model scores, namely, categorization I and II, where in categorization I, “Negative” includes {1,2}, “Neutron” represents 3, and “Positive” includes {4,5}. Later in the study of HSBC models, a new categorization will be introduced, moreover, the difference between model performance on different categorizations will be discussed then.

2.1. Non-deep-learning Methods

Logistic Regression, Naïve Bayes, and SVM are classical machine learning algorithms, which are good for binary classifications, but they can be used as multi-class classifiers as well. For example, D. Phuc and N. T. K. Phung have used Naïve Bayes for classifying messages [7], T. Pranckevičius and V. Marcinkevičius have used Logistic Regression for multi-text classification [8] and Koby

Crammer and Yoram Singer have used SVM for multi-class problem [9]. Accuracies of three models concerning categorizations I and II are given in Table 1 and Table 2 respectively.

Table 1. Summary of Scores of Models wrt. the Categorization I

Model	Score on Training Dataset	Score on Testing Dataset
Logistic Regression	86.7/100	74.8/100
Naïve Bayes	71.9/100	64.2/100
SVM	96.4/100	73.6/100

Table 2. Summary of Scores of Models wrt. the Categorization II

Model	Score on Training Dataset	Score on Testing Dataset
Logistic Regression	82.6/100	49.6/100
Naïve Bayes	61.1/100	40.5/100
SVM	94.7/100	48.7/100

Besides model scores, detailed accuracies of classification on each tag in categorization I are shown in Table 3, and that of categorization II will be presented together with HSBC models later.

Table 3. Summary of Accuracies wrt. the Categorization I

Model	Negative	Neutron	Positive
Logistic Regression	78.36%	16.08%	91.82%
Naïve Bayes	44.03%	0%	99.67%
SVM	75.37%	24.62%	88.65%

2.2. Deep Learning Methods

Deep learning is usually referred to as a general frame of machine learning algorithms adopting neural networks. Two special types, Convolutional Neural Network, and Recurrent Neural Network (LSTM) are studied in this paper. As universally used models, several attempts of using neural networks in NLP already exist. For instance, I M Rabbimov and S S Kobilov have used a general machine learning algorithm for multi-class text classification [10], Yoon Kim used CNN for Sentence Classification [11], Francesco G., Stefano S., and Mario C. have used Deep CNN for Extreme Multi-label Text Classification [12], Pengfei L., Xipeng and Qiu Xuanjing Huang have used RNN for multi-class text classification [13], Jenq-Haur W. have used LSTM for sentiment classification [14], Y. Luan and S. Lin have combined CNN and LSTM for text classification [15]. A summary of the scores of CNN and LSTM can be seen in Table 4.

Table 4. Comparison of Model Scores

Model	Score
CNN	42.0/100
LSTM	44.7/100

2.3. Hybrid Sequential Binary Classification (HSBC)

In this section, we discuss how the binary classification method can be applied to address multi-class classification problems. In general, there are two ways which are called “one-vs-one” and “one-vs-rest” respectively. The difference is the method of splitting the dataset. Suppose we have N classes. For “one-vs-one”, we pick two classes once a time, and train a binary classifier on these two classes only. This procedure is repeated for all combinations of two classes and N(N-1)/2 classifiers are needed in total. While testing, a comparison is applied among these classifiers. For “one-vs-rest”, we choose a class at a time and treat the rest classes as another class, next, we train a binary classifier concerning these two classes. Totally, N classifiers are required. While testing, for each data point, the strategy is to classify it into the class where the highest score is attained among N classifiers.

Nevertheless, we do not simply use these two methods separately but somehow combine them and introduce a new method called Hybrid Sequential Binary Classification (HSBC). The idea is to break down the multi-class classification problem into several binary classification tasks concerning different categorizations where the different binary classifier is applied. In our case, we study the combination of Naïve Bayes, Logistic Regression, and SVM. Since our label space can be naturally oriented, there are natural ways to assign some particular meaning to subsets of label space, we focus on two meaningful categorizations as follows.

2.3.1. Global Sentiment Analysis with Local Refinement

This is essentially the same as the categorization I defined earlier. We apply general sentiment analysis in the first stage. In other words, we assign the tag Negative to the subset consisting of comments of 0 and 1 star, “Neutron” to the subset of comments with 3 stars, and “Positive” to the rest. Furthermore, we combine “Negative” and “Positive” together as “Other” so that the task for our first binary classifier is to classify the data into two categories, “Neutron” and ”Other”. Second, we use a binary classifier to distinguish “Negative” and “Positive” in the subset tagged “Other”. The other two classifiers verify the refinement of the category “Negative” and “Positive” respectively, namely, one to classify in two categories 0 and 1 in “Negative” and the other for classification into 4 and 5 in “Positive”. Thus we need four classifiers in total. The model concerning this categorization is denoted by HSBC1 and the procedure is given in Figure 3.

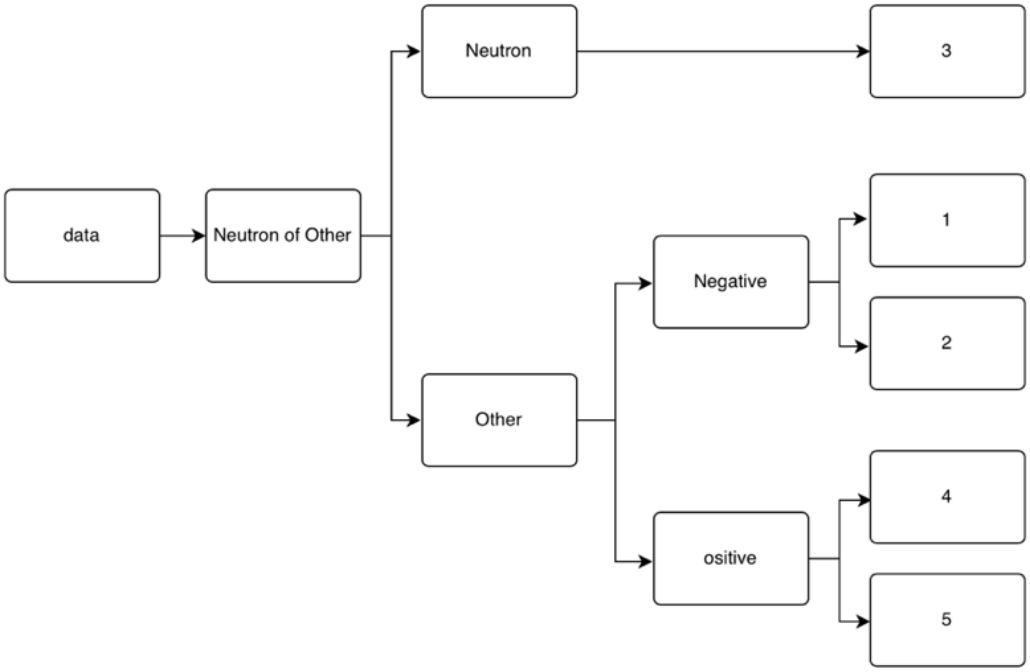


Figure 3. Procedure of HSBC1

“One-vs-rest” strategy is used in the first step, and the rest follows the “one-vs-one” method, thus our HSBC1 can be seen as a model of the half-breed type.

2.3.2. Descend of Extreme Analysis

Instead of objective sentiment analysis, we focus on subjective star evaluations. Particularly, we view the subset of comments consisting of 0 and 5 stars as the subset tagged “Extreme” meaning they are the worst and the best respectively, also, 3 stars present “Neutron” as usual and the subset made up of 2 and 4 stars will be categorized as “Moderate”. In this case, the first classifier determines if the data belongs to “Neutron” or not, the second distinguishes “Extreme” and “Moderate” and the other two refine “Extreme” and “Moderate” respectively. The model concerning this categorization is called HSBC2 and its procedure is shown in Figure 4.

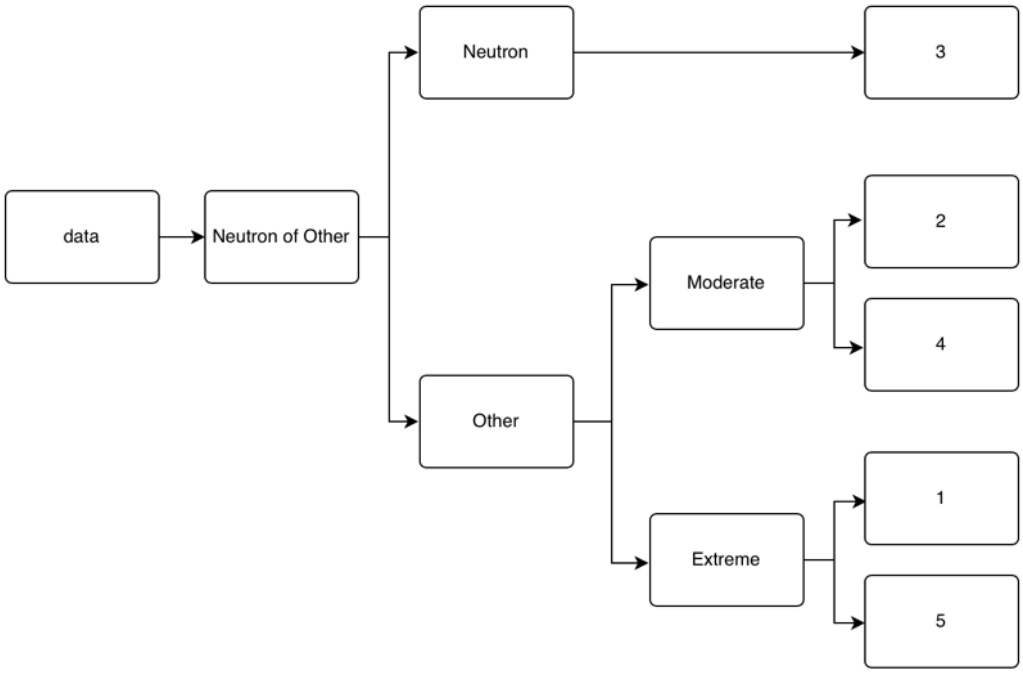


Figure 4. Procedure of HSBC2

For two HCBS models, our strategy is to find the best classifier in each step and glue them together. Moreover, we will compare their accuracies.

We study HSBC1 in more detail. As we have seen above, the first step is to predict if the label of a data point is "Neutron" or not, so we have to explore our three binary classifiers on their performance of classifying "Neutron" and "Other" and pick the classifier whose performance is the best among them.

Table 5. Accuracies on Classifying Neutron and Other

Model	Neutron	Other
Logistic Regression	2.29%	99.61%
Naïve Bayes	0%	100%
SVM	16.57%	95.51%

As shown in Table 5, the classification accuracy of Naïve Bayes on "Other" is the best while that of predicting "Neutron" is the worst. Somehow Naïve Bayes should be thought of as the best classifier we need, indeed, if x whose true tag is "Neutron" is classified into "Other", we convert it to "Neutron" manually, this modification will make Naïve Bayes performs perfectly on classifying "Neutron". After that, we have to choose the best classifier for predicting "Positive" and "Negative". For the rest classifier selection, we will refer to the model scores rather than accuracies on each tag. The accuracy comparison of the rest step is summarized in Table 6.

Table 6. Comparison of Model Scores

Model	Positive/Negative	1/2	4/5
Logistic Regression	86.6/100	65.3/100	71.8/100
Naïve Bayes	76.7/100	62.3/100	70.3/100
SVM	86.7/100	62.5/100	69.5/100

We choose Logistic Regression to classify {1,2}, and for classification {4,5} and "Positive/Negative", we try them all because their performance is tantamount to each other. Finally, we have packaged our HSBC1 model, and the complete procedure is shown in Figure 5.

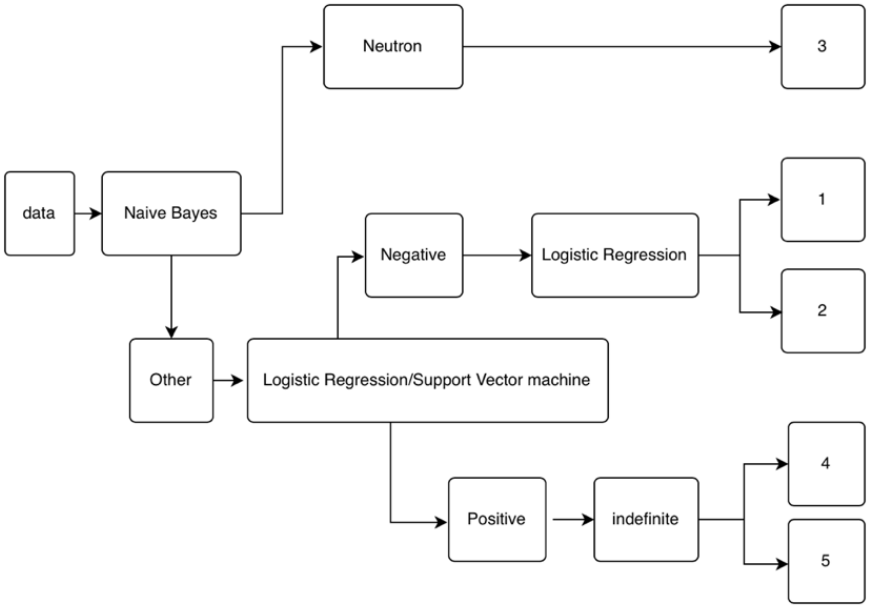


Figure 5. Complete Procedure of HSBC1

As shown in Figure 5 above, there are six possible combinations of HSBC1 and a complete summary of model scores are given in Table 7.

Table 7. Summary of Scores of HSBC1 Models

Model	Positive/Negative	1/2	4/5	Score
HSBC1-A	Logistic Regression	Logistic Regression	Naïve Bayes	63.7/100
HSBC1-B	Logistic Regression	Logistic Regression	Logistic Regression	64.7/100
HSBC1-C	Logistic Regression	Logistic Regression	SVM	61.6/100
HSBC1-D	SVM	Logistic Regression	Naïve Bayes	61.8/100
HSBC1-E	SVM	Logistic Regression	Logistic Regression	62.5/100
HSBC1-F	SVM	Logistic Regression	SVM	61.5/100

The confusion matrix of the best one (HSBC1-B) is given in Figure 6.

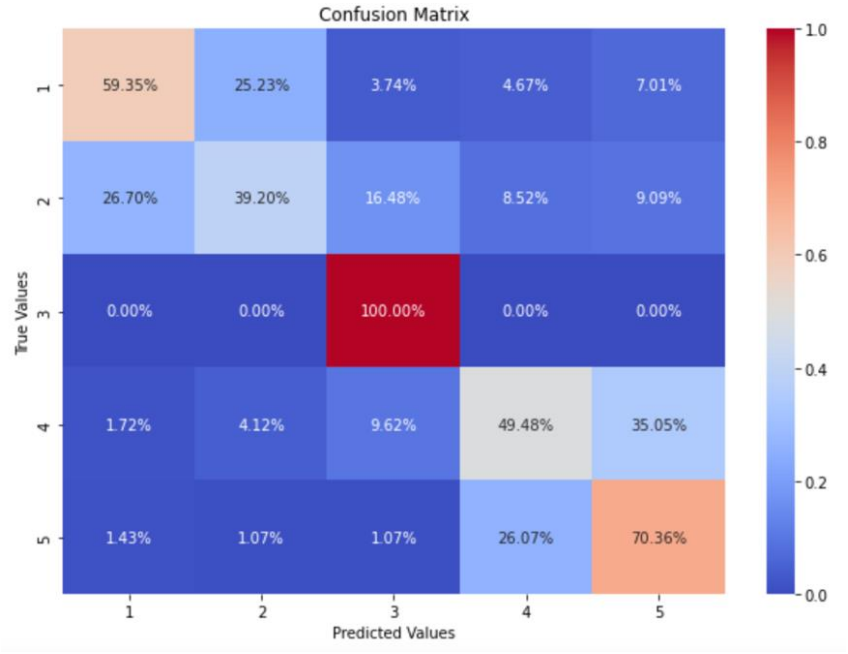


Figure 6. Confusion Matrix of HSBC1-B

We study the second case HSBC2 in more detail. The first step is the same, so we just choose Naïve Bayes as we do in the first case. For the other three classifiers, we also compare their scores, which are summarized in Table 8.

Table 8. Comparison of Model Scores

Model	Extreme/Moderate	1/5	2/4
Logistic Regression	66.2/100	90.2/100	77.2/100
Naïve Bayes	63.4/100	84.4/100	64.0/100
SVM	62.7/100	91.2/100	81.0/100

The second classifier is chosen to be Logistic Regression and the last one is SVM, but for the third one, Logistic Regression and SVM will be both considered. The whole procedure is shown in Figure 7.

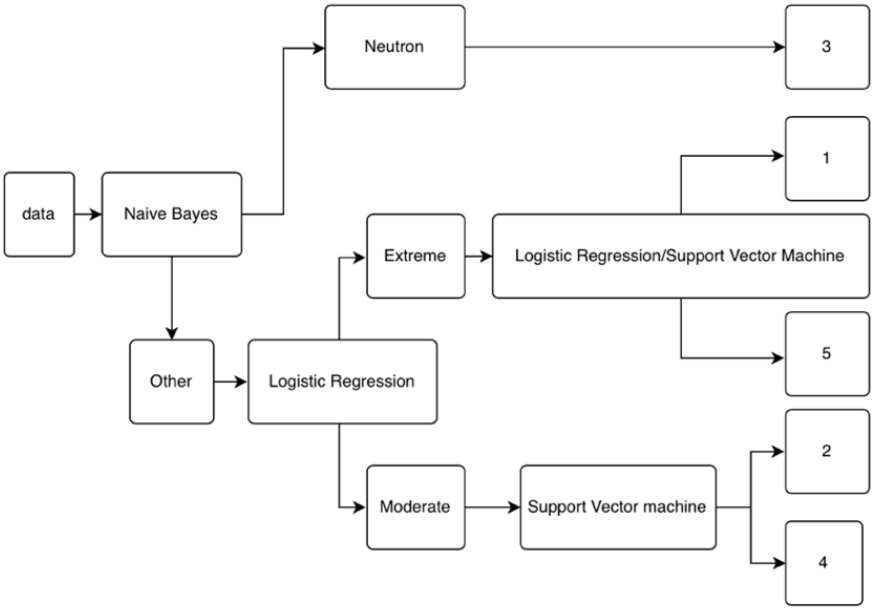


Figure 7. Complete Procedure of HSBC2

Similarly, Table 9 presents the comparison of scores between HSBC2-A and HSBC2-B.

Table 9. Summary of Model Scores

Model	1/5	Score
HSBC2-A	Logistic Regression	62.6/100
HSBC2-B	SVM	60.1/100

3. Discussion

As we have seen in methodology, if we simply apply non-deep-learning methods to our five-class classification problem, they do not behave well, nor do deep learning methods. Classifying “Neutron” is hard for all models we consider, for, neither of which has accuracy over 50%. However, non-deep learning methods have different accuracies concerning different categorizations. Indeed, their accuracies are improved using categorization II. The reason for the bad behavior of classical methods is clear, for they are often used for binary classification, while there may be mainly two reasons for deep learning methods. One is that we do not use so many hidden layers and the other is the data in our case study may be complex. Also, it is difficult to control the number of parameters to get rid of overfitting, which may cause CNN and LSTM to have low scores. For them, the more dense layers do increase accuracy on the training dataset significantly, particularly, the LSTM is more sensitive to the addition of hidden layers, but their accuracies on the testing dataset remain similar. Models of

HSBC type perform better than each component classifier separately, which is a good try for simple usage of binary classifiers to tackle multi-class classification problems. The reason is intuitive, for, we keep the advantage of each classifier and get rid of their shortage by combining them in the way we design. It should be noted that local optimum does not imply global optimum, in other words, combining the best classifier in each step does not mean their combination will be the best. Also, though HSBC1-B behaves the best among all these combinations, the improved performance is the cost of time, for instance, we need more time for finding the "best" classifier on each step, also, as the model is constructed manually, we have to calculate manually its confusion matrix and accuracy respectively ourselves. A complete and detailed comparison of accuracies of all models is given in the following Table 10.

Table 10. Complete Summary of Accuracies wrt Categorization II

Model	1	2	3	4	5
Naïve Bayes	21.96%	7.39%	0.42%	76.63%	52.86%
LogisticRegression	59.35%	39.20%	26.36%	49.48%	70.36%
SVM	54.21%	36.93%	31.80%	45.70%	60.00%
CNN	50.00%	45.36%	39.62%	43.38%	54.72%
LSTM	42.16%	43.81%	34.43%	43.38%	55.03%
HSBC1-A	59.81%	39.20%	100%	62.20%	52.50%
HSBC1-B	59.35%	39.20%	100%	49.48%	70.36%
HSBC1-C	59.81%	41.48%	100%	45.02%	60.00%
HSBC1-D	56.54%	35.80%	100%	59.45%	52.14%
HSBC1-E	57.01%	36.36%	100%	46.39%	67.86%
HSBC1-F	56.54%	35.80%	100%	48.45%	62.14%
HSBC2-A	63.08%	31.82%	100%	39.86%	73.21%
HSBC2-B	54.21%	36.93%	100%	45.70%	60.00%

HSBC2-A has the best performance 63.08% on classifying 1, LSTM has the best performance 43.81% on classifying 2, HSBC models have the best accuracies 100% on classifying 3, Naïve Bayes has the best performance 76.63% on classifying 4 and HSBC2-A has the best performance 73.21% on classifying 5. Overall, the classification accuracy on 4 and 5 is better than that on 1 and 2.

For all models, we consider in this paper, though their accuracies on the training dataset are satisfied, those on the testing dataset are not good enough because we do not analyze features of any particular sentence, which should be the right way to solve any problem in NLP. If we take into account of consideration we just mentioned, the model accuracy on the testing dataset may be improved further. For example, a potential problem lies in when we deal with stop words, recall that stop words are often removed in NLP algorithms, but there are words such as wasn't, mustn't, etc. The loss of words of this kind may confuse classifiers to distinguish emotional comments. Indeed, these words often have negative emotional implications.

4. Conclusion

Starting with the case study of Amazon product review, several classifiers are applied and their accuracies are compared on the testing dataset which makes up 10% of the total data. Though they behave well on the training dataset, they do not keep on the testing dataset. The best accuracy of models on the training dataset is around 99%, obtained by deep learning methods. Indeed, the more hidden layers are, the more accurate models can be. The best accuracy on the testing dataset is 64.7%, obtained by HSBC1-B. It is worth noting that HSBC models behave even better than deep learning methods on our particular dataset. Though the HSBC models do a relatively good job in this paper, they are not limited to the three classifiers used in this paper, in other words, other classifiers which are particularly good on binary classification can be considered and the model performance may be

improved further. Furthermore, HSBC is not the only way to combine classifiers. It is believed that this naïve HSBC modeling will pave the way for more possibilities.

References

[1] Walaa Medhat, Ahmed Hassan, Hodi Korashy, *Sentiment analysis algorithms, and applications: A survey*, Ain Shams Engineering Journal, Volume 5, Issue 4, 2014, Pages 10931113, ISSN 2090-447.

[2] Vishal A. Kharde, S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications (0975 – 8887) Volume 139 – No.11, April 2016

[3] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake News Detection Using Sentiment Analysis," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-5, doi: 10.1109/IC3.2019.8844880.

[4] Alsafy, Baidaa & Mosad, Zahoor & Mutlag, Wamidh. Multiclass Classification Methods: A Review. International Journal of Advanced Engineering Technology and Innovative Science (IJAETIS), 2020.

[5] Fang, X., Zhan, J. Sentiment analysis using product review data. *Journal of Big Data* **2,** 5 (2015). https://doi.org/10.1186/s40537-015-0015-2.

[6] Abinash Tripathy, Ankit Agrawal, Santanu Kumar Rath, Classification of Sentimental Reviews Using Machine Learning Techniques, Procedia Computer Science, Volume 57,2015, Pages 821-829, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.07.523.

[7] D. Phuc and N. T. K. Phung, "Using Naïve Bayes Model and Natural Language Processing for Classifying Messages on Online Forum," *2007 IEEE International Conference on Research, Innovation, and Vision for the Future*, 2007, pp. 247-252, doi: 10.1109/RIVF.2007.369164.

[8] T. Pranckevičius and V. Marcinkevičius, "Application of Logistic Regression with part-of-the-speech tagging for multi-class text classification," *2016 IEEE 4th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, 2016, pp. 1-5, doi: 10.1109/AIEEE.2016.7821805.

[9] Koby Crammer, Yoram Singer, *On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines*, Journal of Machine Learning Research 2 (2001) 265-292.

[10] I M Rabbimov, S S Kobilov, Multi-Class Text Classification of Uzbek News Articles using Machine Learning, Journal of Physics: Conference Series **1546** (2020) 012097.

[11] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

[12] Francesco Gargiulo, Stefano Silvestri, and Mario Ciampi, Deep Convolution Neural Network for Extreme Multi-label Text Classification, *Institute for High-Performance Computing and Networking, ICAR-CNR, Via Pietro Castellino 111 - 80131, Naples, Italy.*

[13] Pengfei Liu, Xipeng and Qiu Xuanjing Huang, Recurrent Neural Network for Text Classification with Multi-Task Learning, Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16).

[14] Jenq-Haur Wang, An LSTM Approach to Short Text Sentiment Classification with Word Embeddings, The 2018 Conference on Computational Linguistics and Speech Processing ROCLING 2018, pp. 214-223.

[15] Y. Luan and S. Lin, "Research on Text Classification Based on CNN and LSTM," *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 2019, pp. 352-355, doi: 10.1109/ICAICA.2019.8873454.