Applied Machine Learning Report

ILKER OZTURK
APPLIED MACHINE LEARNING
ID: 001187380

Abstract*— Within the scope of this coursework, ML programs have been created that predict which category the product belongs to and how many stars it receives from the reviews of Amazon's customers. The project consists of 2 parts, in the first part there are Exploratory Data Analysis and Data cleaning sections that I also used in the later sections. In first part, the problems in the existing data are identified and if possible, this data is extracted if it is not transformed into a useful structure using managing missing data methods. After that, in the first part, encoding target and feature scaling parts are implemented and then the performances of basic ML models are observed. Hyper tuning is applied to the algorithm that showed the best performance and its performance is increased. After that, Neural Network (NN) algorithms are tested, and performance comparisons are made. In the basic ML methods, Logistic Regression, SVC, Decision Tree, KNN and Naive Bayes methods were tested in both parts. In the first part of the project, Convolutional Neural Networks (CNN) is used as the Neural Network algorithm. In the second part, Recurrent Neural Network-LSTM (RNN) is used as the Neural Network algorithm.

I. Introduction and related work

This section should talk about the following: This project aims to estimate the category of the product and the number of stars that customers give to the products, based on the comments made by the customers. The most important thing here is to find the connection between the comments written by the customers and their buying behaviour. The most important challenge is that people sometimes make sarcastic comments and say the opposite of what they actually want to say. Since sufficient data is provided, correct inferences can be made by learning this by the algorithm. The important thing in this regard is the quality and adequacy of the data. The important thing for this issue is to analyse the data well, to correct the corrupted parts and thus to increase the quality of the data. This is a problem that needs to be solved in the most effective way, especially by all commercial companies. In this way, it is ensured that the customer-related behaviour analysis of the shopping behaviours of the customers is made, and the products that meet the customer's requests are produced or recommended to increase the sales and profitability. Today, similar studies are also carried out by academics and technology companies. Since some of them are within the scope of trade secrets, companies can keep them. There are many similarities to the ML research conducted here. According to (Naseem et al., 2021), “Various perspectives on classification systems were categorised, along with their benefits and drawbacks. Due to their simplicity and high degree of accuracy, supervised machine learning algorithms are usually the most extensively utilised methods in this discipline. Classification based on naive Bayesian and SVM algorithms, which are commonly referred to as baseline methods for comparing approaches to freshly suggested methods. In terms of accuracy of results, the CNN model was more efficient for deep learning.” The results of this research have similar results with the text classification research I did on amazon reviews. In this data set, there are review id, text, verified product category and review score fields. Incorrect data was detected in this data set. It has been determined that there is an erroneous data such as “-1” in the review score field, which should be between 1 and 5. Using missing data techniques to handle this data, the loss of valuable data is prevented. The weight of 901 lines with “-1” value is around 3% compared to the total. Since the review stars value is over 4 and the data in the text is generally 5 weighted, these data have been changed with the most frequent value. In the current data structure, 14 rows contained empty values. These 14 rows have been dropped because there is no chance of recovering this data. These rows are converted to csv file and included in the coursework report. On the data scaling side, SMOTE was used and this technique was also used for the second problem as the results were close to perfect for the first problem. SMOTE is a very popular oversampling technique today and effectively solves oversampling problem by generating synthetic data. In such applications, some words can be dominant because they are used too much. This problem is solved by using TF-IDF. In the ML part, CNN and RNN(LSTM) using word embedding. The reason why I use CNN is because it is accepted as a very successful algorithm in classification, and it is an effective algorithm that detects the properties of texts with its different layers and can classify them. The reason why I use RNN is because it is an ML algorithm that progresses by remembering the previous data. LSTM, on the other hand, is a subclass of RNN and provides long-term recall of data. It makes sense to use RNN as a text classification method, since the words that come one after another in the daily language relate to each other. Within the scope of Coursework, first the basic ML methods were tried to determine which one was more effective and then their performance was improved by using hyper tuning, and then the same problem was tried to be solved with the neural network algorithm. Then, the performance of the algorithm is tested with random data selected from the test data.

II. Ethical discussion

There may be people and situations that each ML algorithm negatively affects. The customer shopping behaviour analysis included in this project can lead to major problems if the

privacy of personal information is violated. From this point of view, since the private information of the customer is processed within certain rules and in the ML algorithm, only customer behaviour data should be included, not this information. In addition, it should not be forgotten that the data can be evaluated within the scope of legal rules and within the scope of the permissions given by the customer in this direction. Understanding the customer's opinions and comments about a product and what he or she thinks about which product, suitable products can be recommended for them, but this can cause very distressing results in some circumstances. Since the ML algorithm does not have moral values such as discrimination of social groups and gender equality, the results may cause discrimination for some segments of the society. For example, if this dataset had an age column, gender column or location information, it would be unethical to only recommend cheap products to some customers or recommend cleaning products for a specific gender reason. In general, some limiters are defined against such discriminations, and it is aimed to avoid distinctions such as gender, religion, ethnicity. Although there is no such data in the current dataset, datasets that can categorize in a large dataset in the future should be avoided or this should be prevented with open-source code restrictors that provide these restrictions today. While designing ML algorithms, the social problems that may arise in this way should be proceeded by taking into account each step.

III. Dataset preparation

First, the content of the data was analysed in all details with the help of visuals. Then, anomalies in the data were detected. In the review star field, which should have values between 1 and 5 in this data set, a value that should not be like "-1" was detected in 907 lines. When the data is examined, it is observed that the texts here are of good quality and these values were entered as a result of an error, but these data generally contain texts compatible with 5 stars. For this reason, when the data is so precious, the missing parts are filled with the most frequent technique, which is a method of handling missing data. Even if the average value technique is used here, the result will still be the same as the average is over 4 units. Also, 907 lines are only 3% of all data, so it won't make a dramatic difference. For these reasons, it is thought that data recovery is a more appropriate option instead of deleting the existing data. Finally, it has been determined that the text fields in 14 lines are completely empty. Since the presence of blank data in this field will adversely affect the performance of the neural network, this data has been extracted and saved in a CSV file. Afterwards, stemming method was used since NLP is generally focused on the root of words. I decided to use PorterStemmer() because the it is the most common stemming way in English language. In neural network solutions, on the other hand, similar words are presented in a similar way by using the Word Embedding method. After that, With the TF-IDF (Term Frequency - Inver4se Document Frequency) method, it is ensured that the statistical significance of the terms in the reviews is calculated effectively. Then, the SMOTE technique was used to make the unbalanced values balanced so that the algorithm would be effective. With this oversampling technique, synthetic data has been created and the data set has been made balanced. Today, this method is the most popular scaling method, and it has been preferred in the second part because it gives excellent results in solving the first problem.

IV. Methods

While solving both problems, I first determined how effective the basic ML algorithms were in solving the problem. I compared the performances of Logistic Regression, Decision Trees, KNN, Support Vector Machine and Naive Bayes. These algorithms are classification algorithms that are generally used for classification and have proven to be very effective in some cases and are used to make predictions. The reason for using CNN is because it is a classification algorithm such as logistic regression as one of the two basic algorithms of the neural network and it is well-suited for this problem. The algorithm consists of 3 main parts, namely input, hidden layers and classification, and it progresses by providing step-by-step learning due to the structure of the algorithm. The algorithm providing the most efficient result for this problem was CNN with a rate of 92%. Then, the RNN algorithm, which is another basic algorithm of the neural network, is preferred. RNN is an algorithm that proceeds by learning briefly from the previous steps. The reason why the LSTM algorithm, which is a subset of RNN, is preferred is to proceed by determining what will be forgotten and remembered with the GATE structure in the LSTM algorithm. LSTM is used in many topics such as sentiment analysis, text generation, and time series. The reason why it is also used for this problem is that it is thought that solving the consecutive relationship between words in daily language will solve the problem more effectively.

V. Experiments and evaluation

First, basic ML methods are decided in the first part. The highest performance values belong to the Logistic Regression and Support Vector machine algorithms. Afterwards, the performance of the Logistic Regression algorithm was tried to be increased by applying hyper tuning. By using GridSearch(), the values giving the highest performance were determined (87%). The performance of Logistic Regression has been increased to 90%. This rate is quite high. Because under normal conditions, the product category estimate is 50% under normal conditions. Afterwards, a basic CNN algorithm was created, and then layers were added sequentially. In the last case, with the Embedding layer, a structure with the GlobalMaxPool1D layer, the output layer with the relu activation function, and the regularization with the dropout layers is obtained. Here, "adam" is used as the optimizer and categorical cross entropy is used as the loss function. The highest result with 92% belongs to the CNN algorithm. In addition, it is shown how the optimized performance of logistic regression is reduced by changing the best parameter values in the coursework code document. The performance values of the algorithms for the first part are as in the table.

Algorithm	Performance
CNN	92%
Logistic Regression (Hyper Tuned)	90%
Logistic Regression	87%
SVC	87%
Decision Tree	82%
Naive Bayes	67%
KNN	61%

Table 1:Product Category Classification Performance

In the second part, the performance of estimating the number of stars between 1 and 5 given to a product was calculated. Here, first the basic ML algorithms and then the RNN were used to improve the result. Optimized Logistic Regression provided the best performance. The highest parameter values of the basic algorithm were found again with GridSearch(). Afterwards, the neural network solution of the problem was provided with RNN. In this algorithm, Adaptive algorithms generally outperform the default SGD in terms of speed. For this reason, the Adam optimizer is used. Afterwards, the "softmax" value in the activation parameter of the RNN algorithm was changed with "relu", and its performance was observed, and a 12-fold worse result was obtained. This is because "softmax" is a function for multiple classification problems, and relu is a nonlinear function. In this section, Logistic regression and SVC were shown as the best performance values, and since it took a long time to find the optimized parameters of SVC with GridSearch(), the best parameter values of Logistic Regression were found. Both algorithms have been tested with 3-10 test data in the coursework report. In this problem, in fact, a performance gain from 20% to approximately 70% has been achieved. Optimized Logical regression and RNN gave almost the same results. In addition, it was observed that in the tests performed with the data selected manually from the test data, it was observed that the rate of fully correct prediction was high, and in the wrong predictions, predictions very close to the real value were made, and it was observed that the RNN algorithm worked effectively. The performances of the algorithms are expressed in the table below.

Algorithm	Performance
RNN	67%
Logistic Regression (Hyper Tuned)	69%
Logistic Regression	66%
SVC	66%
Decision Tree	58%
Naive Bayes	57%
KNN	59%

Table 2:Review Star Prediction Performance

VI. Discussion and future work

Reflections on a) what worked well and what worked less well; b) reasons behind the performance obtained; c) how your work could be extended in the future and what addition can be made to it. In general, it is observed that a high-performance gain is achieved. The probability of knowing correctly, which is 50% in the first part, has been increased to 92% with the neural network. In the second part, it was observed that this rate was increased from 20% to 70%. In the observations made, although it performed well with logical Regression in SVC, logical regression best parameters were found in both cases because it took a long time to be optimized. Since the sent code will be compiled again, the best parameters were not found using SVC GridSearch() to avoid a time problem. This can also be tested in future developments. The most important problem in this dataset is that sarcastic interpretations mislead machine learning. This problem may be solved in the future by adding a line to the dataset that includes how useful the comment was made by other users. In this way, results can be improved by increasing the weight of more useful comments in the analysis rather than useless and misleading comments.

VII. Conclusions

The category and review score values of the product were estimated by using the reviews in the data set of the Amazon company within the scope of Coursework. First of all, the data was extracted with the steps of Exploratory Data Analysis and Data cleaning. Then, it was ensured that the data became balanced from an unbalanced state and the text structure was changed according to the machine learning method and made it ready for the ML algorithm. Then, the results were compared using basic ML algorithms and the best algorithms were optimized by finding the best parameter values. Then, a solution to this problem was provided with the neural network algorithms CNN and RNN. The results were observed by testing the code with optimization and test data. Overall, the results were very good, although suggestions were also made on how the results could be improved in the future.

REFERENCES

[1] S. Naseem, T. Mahmood, M. Asif, J. Rashid, M. Umair and M. Shah, "Survey on Sentiment Analysis of User Reviews," *2021 International Conference on Innovative Computing (ICIC)*, 2021, pp. 1-6, doi: 10.1109/ICIC53490.2021.9693029.

[2] Openreview, github. Accessed: 01/04/2022. [Online]. Available: https://github.com/aniass/Product-Categorization-NLP/

[3] Openreview, github. Accessed: 01/04/2022. [Online]. Available: https://github.com/BenRoshan100

[4] Openreview, github. Accessed: 01/04/2022. [Online]. Available: https://github.com/susanli2016/NLP-with-Python