Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset

Mujiono Sadikin1,* and Abi Fauzan2

ABSTRACT

Due to the abundance of text data representing public opinion, the Sentiment Analysis study is getting more and more important. Various techniques and methods have been proposed to address the issues. One of those techniques is deep learning algorithms which have been used to achieve great results in Natural Language Processing (NLP) applications. Sentiment Analysis is a part of NLP application that extracts emotional information from texts. In this study, we investigate the performance of sequence-based model, i.e., LSTM, compared with multi-layer perceptron Neural Network (NN) to classify the polarity of the text review based on negative or positive. The dataset used in this study is a restaurant review taken from the Yelp website. The dataset is trained using Word2vec word embedding to convert words contained in the dataset into numerical vector representation which is used as the deep learning model input. Based on the experiment results, it is shown that the LSTM model is outperformed compared to the multi-layer NN model. The best accuracy performance provided by LSTM model is 91%, whereas the best accuracy performance of multi-layer NN model is 76%.

Keywords: NLP, yelp dataset, sentiment analysis, word embeddings.

1. INTRODUCTION

Billions of data, including text data, are generated from various sources on the internet every day. Text-based data can be found on social media services, forums, articles, and sites containing public opinion on products, brands, services, politics, or any other topic. Some of the data utilizations by companies, governments, and other fields are to create marketing analyzes, product reviews, or community services. That’s why the test data as input from users or consumers that is efficiently analyzed is very valuable. The analysis process can be carried out by using the Sentiment Analysis (SA) approach. SA is one of the opinion mining studies as a process of understanding and transforming unstructured data into structured data automatically to get an understanding of the information contained in an opinion sentence [1]. SA is able to assist in determining opinion as positive or negative from the text towards which sentiment is directed [2]. As the task of Natural Language Processing, the usage of SA approaches cover a wide range of area such as drug review [3], entertainment [3], [5], hospitality & tourism [6], and covid-outbreak [7], [8].

The essence of SA system is finding the meaning in a single sentence, it is occasionally containing some opinions or facts in unstructured conditions [9]. Therefore, it is required to transform this kind of sentence into a structured condition. One of the necessary processes in the SA system is to extract words into a numerical representation namely feature extraction, in which computers can process it. In general, feature extraction techniques are Bag of Word, TF-IDF, and Word Embedding (Word2vec, GloVe) [10], [11]. These techniques are useful to explore the potential information on the datasets or dictionary and represent original samples as feature vectors which are used as input to the machine learning techniques. Word embedding with Word2vec technique is used in this research.

Many researchers have presented the effectiveness of deep learning techniques in sentiment analysis topic to extract meaningful information expressed by people. Multi Laye Perceptron (MLP) is a type of feedforward deep learning network. It has achieved great results in sentiment analysis tasks. Another type of deep learning network that can solve sentiment analysis tasks is LSTM. Text review is a sequence data, LSTM network takes advantages of this problem since it learns context in chronological order from each text in dataset. In addition, the case of a long
review text, thanks to additional layer in LSTM cell namely memory cell which will decide what information will be passed or removed.

As an open dataset, Yelp database has attracted many researchers to investigate various solutions to address the challenges offered. One of studies by using Yelp dataset is conducted by Fan and Khademi in [12] on the restaurant’s average star rating prediction. The authors utilized the unigram model combined with Part-of-Speech (POS) tagging as a feature engineering, Support Vector Machine (SVM), and decision tree were applied for prediction. The dataset used in their experiment consists of 4234 lists of restaurants and 35465 reviews. Their study relates to this proposed research in the sense of NLP task which is opinion mining. In terms of the feature extraction process, the implementation of POS tagging has indeed produced a good feature vector, but this does not include the semantic meaning of the data. Thus, compared to the proposed research, the deep learning models are associated with Word2vec word embedding which generates a feature vector with meaningful representation.

The superiority of LSTM in handling SA task for sequence dataset is shown by Miedema et al. through their publication [13]. In the study on SA task, authors proposed an advanced study used to classify tweet data. They used LSTM model to do classification tasks which are often used in long-term sequence data. The result of their research achieved the accuracy of 86% correctly classified. There was no pre-processing phase as they used a pre-build IMBD dataset that was already vectorized in number format. Furthermore, performing hyperparameter optimization on their model parameters could improve the result of accuracy.

A comparison of classification methods in extracting drug name entities in the medical text is proposed by Sadikin et al. in [14]. They proposed a new representation technique based on the characteristics of word distribution and word similarities to overcome some issues in medical text mining, such as multiple names for the same drug, the lack of labeled dataset sources in medical domain, and external knowledge. In their research, three data representation techniques were conducted, the first was with a standard neural network model, that is MLP, and the second techniques were Deep Believe Network (DBN) and Sparse Autoencoder (SAE). The last technique was using LSTM model. The results of all models have given the best result to LSTM model compared to the state-of-art, with its average f-score being 86%. The study results convince that the LSTM model gives the best performance for the sequence data representation as the representation of medical text sentences. Another study utilize the LSTM model is conducted by Wiranda et al. in [15]. In the study, authors use the LSTM technique to predict product sales in the pharmacy industry. The main topic of their research is the ability to predict the direction of prices of the company’s product sales. In their research, authors treat the dataset as time-series data. As the task to be handled is to predict the continuing number, the performance parameters evaluation used are Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). The experiment results show that the LSTM network model achieved the best performance compared to the other models.

An experimental study on sentiment analysis by exploring various deep learning based classification models is conducted by Colón-ruiz and Segura-bedmar as published in [3]. In their study, authors compare and analyze some of the deep learning models i.e., tf-idf SVM, word embedding-CNN, and word embedding with the combination of CNN and LSTM. Of those models explored, the study reveals that Bidirectional LSTM with BERT word embedding is a superior configuration to address the challenges in drug reviews from medical corpus published in drug.com.

Rather than using the English text as the mainstream dataset for SA, Dashtipour et al. proposed another language text dataset to perform SA study, i.e., Persian Language [4]. The Persian text dataset contains a review of Persian Movie. However, the study also shows that LSTM model applied to non-English text is superior compared to MLP, SVM, and CNN as well. The study also convinces that all variation of LSTM approach achieves better accuracy performance compared to the other models studied.

LSTM as the most suitable model to attack the SA problem of text dataset set is also confirmed by Martin et al. as their study results presented in [6]. The authors used the tourism area as their case study by using the customers’ reviews of booking.com and tripadvisor.com. In the pre-processing stage, the authors utilize NLTK library to perform the pre-processing tasks such as tokenization, stop word removal, and data representation which is ready to be consumed by deep learning models. In the model training and testing authors explore the performance of CNN, LSTM, and the combination of both models to classify each instant data into “Good” or “Bad” classes. The architecture of LSTM-300 presents the best accuracy of 89.19% whereas CNN-LSTM is 88.3% and CNN-64 is 88.2%.

Based on some literature reviews regarding Yelp dataset and LSTM above, in the study, we investigate the application of LSTM models to the Yelp text dataset containing restaurant-business customer’s reviews. As the comparison model, we used the MLP configuration. By modifying the batch size of those two models, as the results of the final experiment, it is shown that LSTM model with a batch size of 32 provides the best accuracy performance. In this proposed study, SA is performed to classify the polarity of given restaurant reviews whether it is a positive review or negative review which is known as a binary text classification task. The dataset used in this research is Yelp dataset which contains more than 6 million restaurant reviews in the USA. Each instant of a dataset has 9 attributes i.e., review id, user id, business id, stars, data, text, useful, funny, and cool [16]. The attributes used for this research are only text (contain a review from a user) and stars (contain a rating of the restaurant). Due to high computational process and education purposes, the dataset is truncated to only 100,000 reviews which contain 50,000 score ratings above 3 and 50,000 score ratings below 3. Feature vectorization is performed by using Word2vec word embedding technique [17]. In the experiment, Recurrent Neural Network (RNN) particularly LSTM network
Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset
Sadikin and Fauzan

is implemented as a deep learning with hyperparameters optimization such as differentiation of batch number and the implementation of dropout regularization for validating the generated models.

The research question in this study is how to propose effective deep learning models that can make a prediction on a given new restaurant review data. To address this question, a comparison of MLP model and LSTM model associated with Word2vec word embedding is conducted. The study result showed that LSTM model outperformed the MLP model. The strong idea of LSTM model is that it allows to control of the text flow over the long text. For validation process, both models are compared with different hyperparameter configuration to conclude the research study. Splitting the dataset into a train set and test set is the major idea to validate the result. Based on the experiment results evaluation, the LSTM model provides the best performance which shown by its accuracy value is 91%.

2. Material and Method

This study aims to analyze the reviews and classify the sentiment based on the reviews available. The methodology used in this study consists of five main steps: data collection, data pre-processing, sentence conversion, classification model, and evaluation. The research stage is described in Fig. 1.

The first process of the research stage is data collection and data exploration. The restaurant review data was taken from the Yelp website, www.yelp.com/dataset. A binary classification is performed where each review contains positive or negative sentiment [2]. Data pre-processing includes stop word and punctuation removal, case folding, and stemming. The next process is the conversion of the cleaned dataset into data that is ready to be used as the input for the classification model [18]. This process includes making a word dictionary from the dataset and converting it into numerical vectors. Word to vector conversion is conducted in which the vector values are taken from the training dataset of the restaurant review by using Word2vec word embedding. MLP and sequential model LSTM are then applied to recognize the sentiment analysis of the restaurant reviews. LSTM network is an advanced version of RNN that is mostly used to process long sequential data such as SA [19]. The final process is to evaluate the accuracy of the model for restaurant reviews. The evaluation step is carried out by the computation of performance parameters sourced from the confusion matrix.

2.1. Data Collection

In this study, a restaurant review of Yelp Dataset is used. Yelp Dataset is a collection of user data, reviews, and businesses in a large-scale dataset provided by Yelp company used for individual, educational and academic purposes [12]. The review datasets consist of more than six million reviews of data generated by the users. Due to the high computational deep learning training process, the review data is reduced to 100,000 reviews data which consists of attributes text (review data) and stars (rating of the restaurant). Each review is converted into binary labels 0 and 1 where rating 1–2 grades as negative and 4–5 grades as positive. Table I shows a sample of data review which will be processed in a further step.

<table>
<thead>
<tr>
<th>Text</th>
<th>Stars</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>The best in Wisconsin and being the dairy state that saying a lot the ice cream is a great consistency and there are so many creative flavors like cake batter fudge zanziberry shortcake…</td>
<td>5</td>
<td>pos</td>
</tr>
<tr>
<td>Worst experience ever! Completely unorganized, unfriendly staff and the food was cold and bland,... extremely disappointing</td>
<td>1</td>
<td>neg</td>
</tr>
</tbody>
</table>

2.2. Data Pre-Processing

Pre-processing is carried out by using the scikit-learn’s train_test_split function as the first process to split the review data into 80% train set and a 20% test set. To embed the dataset into the Word2vec model, the data is processed with a function that returns a list of parsed sentences with integer values. For each sentence, noise removal is performed by removing punctuation, numbers, and HTML tag, stop word removal, and case folding. The pre-processing stages in its results sample are presented in Table II.

2.3. Sentence Conversion

Sentence conversion is processed using Word2vec model after the pre-processing step is finished. Vocabulary is built by embedding training data into Word2vec model. This vector conversion learns and sets the vector representation of each word of the dataset [20]. By default, the training
algorithm in Word2vec is CBOW. Two algorithms provided by Word2vec model are CBOW and Skip-gram [21].

CBOW algorithm is a process of calculating the current word from its context by maximizing the probability of the target word by considering its environment context. On the other side, Skip-gram algorithm is a process of calculating the context of word from the word at the current position in the sentence. The Embedding Word vector will produce a padded list of sentences as classification features. The implementation of using Word2vec as a word embedding technique is able to improve the classification model performance since the distance between word-vector represents the similarity between words [13].

### 2.4. Classification Model

The main idea of the classification model in the study is to train the LSTM network and compare it with MLP model through text review data. The MLP architecture is adopted from [22]. MLP can be described as feedforward NN which does not have a deep layer, thus the information flows in a simple direction. Tokenization proceeds in each review produced a list of parsed sentences and is generated a vocabulary of words. The sentiment classification is performed by feeding the word embedding obtained through a vocabulary index-based into the LSTM network. The proposed LSTM model is illustrated in Fig. 2.

Regularization technique such as Dropout is applied in the LSTM model to prevent the model from being overfitting, it could affect the performance of LSTM model [23].

In the process of training the model, the use of Adaptive Moment Estimation (Adam) optimization as described in [24] is implemented for updating the weights and bias values of each neuron across iterations. To predict the outcomes of sentiment, we apply a sigmoid activation function to scale the output data points into values between 0 and 1 [25].

As depicted in Fig. 3, a flattened layer is applied after the embedding process. This layer reshapes the dimension of the input vector equal to a one-dimension array of elements. Thus, the flattened layer from the input is classified in a Dense layer with two nodes and activated with sigmoid function.

### 2.5. Model Evaluation

To both MLP and LSTM models explored in this study, we apply batch size parameters adjustment scenario. The batch sizes investigated are 32, 64, and 128, whereas the other parameters are fixed as presented in Table III Model Configuration. Based on the Confusion Matrix as described in Table IV, the performance parameters are then computed. Adopted from [26], the performance parameters of the models are computed by formula as shown in (1–3).

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]  

(1)
Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset

Fig. 3. Proposed MLP architecture.

### Table V: MLP Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Epochs</th>
<th>Batch size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>5</td>
<td>32</td>
<td>76%</td>
</tr>
<tr>
<td>Model 2</td>
<td>5</td>
<td>64</td>
<td>75%</td>
</tr>
<tr>
<td>Model 3</td>
<td>5</td>
<td>128</td>
<td>74%</td>
</tr>
</tbody>
</table>

### Table VI: LSTM Model Performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Epochs</th>
<th>Batch size</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>5</td>
<td>32</td>
<td>91%</td>
</tr>
<tr>
<td>Model 2</td>
<td>5</td>
<td>64</td>
<td>90%</td>
</tr>
<tr>
<td>Model 3</td>
<td>5</td>
<td>128</td>
<td>87%</td>
</tr>
</tbody>
</table>

3. Results and Discussion

The experiment is performed by using 100,000 restaurant reviews from Yelp dataset. The dataset consists of 50,000 positive reviews and 50,000 negative reviews as well. Thus, the class distribution is balanced already. We used the random data splitting scenario to split data into training dataset and testing data set. The training set is defined as 80% of the data and the testing set is defined as 20% of the dataset.

The testing performance of MLP model is presented in Table V. Model 1 with a batch size of 32 provides the highest accuracy of 76%. A smaller batch size has shown performs better than a larger batch size, and this will give to a higher accuracy result in neural network. Fig. 4 shows accuracy and loss value of the MLP model 1.

![Fig. 4. MLP model accuracy (a) and loss (b).](image)

According to the design of a multi-layer perceptron, each input acts independently, in other words, each input in MLP model has no relation from others input. Thus, this kind of model does not efficiently perform classification in sequential data. Also, based on the above loss result, the test loss is higher than the training loss at the second iteration which caused overfitting in the neural network model. To overcome this issue, the LSTM model which learns better on sequential data is performed to compare the result of accuracy.

Table VI present the performance of LSTM models fed with the word-embedding vectors as its input. It can be examined that the performance model of LSTM network with Word2vec word embedding with any different batch size parameters is superior compared to the MLP Model. The LSTM model learns better than MLP, this is simply because LSTM cell contains a hidden state called a memory cell that captures context from the input and determines which value to process or eliminate in the next calculation. Among of three LSTM Model, Model 1 with 32 batch size values achieved the best accuracy result of 91%. The detail of the LSTM Model 1 performance

\[
\text{Precision} = \frac{(TP)}{(TP + FP)} \tag{2}
\]

\[
\text{Recall} = \frac{(TP)}{(TP + FN)} \tag{3}
\]
of each epoch in the testing phase is depicted in Fig. 5. Fig. 5 demonstrates that in epoch value of 5, the model achieves the best performance and the performance tends to stagnate when we try to increase the epoch value.

The analysis of LSTM and MLP model performance in the testing stages is based on the confusion matrix as presented in Tables VII and VIII. In MLP model, there are still many incorrect class predictions compared to LSTM model. This issue occurs according to the training process of MLP which uses the basic implementation of a multi-layer perceptron where each input has no relation to previous hidden state or previous input. As shown in both confusion matrix, the incorrect class prediction by LSTM model is reduced by almost 20%.

4. Conclusion and Future Study

The study explores LSTM model to address the challenge of SA contained in unstructured text datasets. To validate the LSTM model, we compare it with the MLP model as baseline. According to the experiment result, it is convinced that the LSTM model outperforms compared to the MLP model, due to the control text flow in the LSTM cell so it can learn the context of input data efficiently. This study also reveals that the LSTM model is suitable to solve the problem of sequential text data classification. Furthermore, the correct adjustment of the parameter values of different batch sizes can improve the performance of the model. Based on the experimental results, using a batch size of 32, the LSTM model can provide the highest accuracy of 91%. In this case, the use of a higher batch value does not guarantee better accuracy performance than the smaller one. Based on the experimental results and analysis, with an accuracy of 91%, the use of the LSTM network combined with Word2vec word insertion and some hyperparameter settings can be used in the case of sentiment classification, especially in restaurant review classification on Yelp dataset. In the future, the proposed model can be implemented in other datasets in determining polarity based on sequence data such as text reviews. The application of stacked LSTM and bidirectional LSTM that has the potential to improve performance also needs to be investigated, as well as other word embedding techniques such as GloVe and Fasttext [27], [28].

**Conflict of Interest**

The authors declare that they do not have any conflict of interest.

**References**


Evaluation of Machine Learning Approach for Sentiment Analysis using Yelp Dataset

Sadikin and Fauzan


