Performance Analysis of Machine Learning and Deep Learning Algorithms for Sentiment Analysis

Mugdha Deokar
Electronics and Telecommunication
Savitribai Phule Pune University

Varun Godse
Electronics and Telecommunication
Savitribai Phule Pune University

ABSTRACT
Public opinion or review of a product, a movie, or a restaurant is a key driver of trends and influences how a person chooses a particular service. This out-pour of opinion contributes to the overall information and becomes a crucial tool to analyze the sentiment towards the service provider. A large pool of information thus can be processed with the help of Natural Language Processing (NLP) tools to find out how good or bad a movie, product, or restaurant is. A dataset containing sentences from websites like Amazon, IMDb, and Yelp to study and predict people's sentiment towards a particular service is used. To achieve the results, Natural Language Processing for sentiment analysis is utilized. Initial step of the project includes designing a model based on six machine learning algorithms like Support Vector Machine, Random Forest, and other algorithms for classification purposes. Next step was using a voting classifier to extract the best features of each algorithm and get a conclusive result. Further, Recurrent Neural Networks and Long Short Term Memory(LSTMs) leverage the power of deep learning to achieve higher accuracy and better results. Also, the BERT model is used to perform sentiment analysis. Thus, the paper aims to compare various possible algorithms that can be used for sentiment analysis using machine learning, deep learning, and natural language processing. In the end, it is investigated if the public response to the service provided is either positive or negative.

General Terms
Machine Learning, Deep Learning, Computer Vision, Sentiment Analysis

Keywords
Sentiment Analysis, Natural Language Processing, Machine Learning, Deep Learning

1. INTRODUCTION
In any service business, the reviews and feedback from the customers are crucial aspects of the business model. This not only helps the service provider but also gives an insight into the quality of the product. In this project, three such services are considered, i.e., restaurants, movies, and e-commerce. These reviews are often positive or negative and have an impact on how popular the service gets and helps settle a score among different choices available to the customer. When the decision is of high value (Buying an expensive product/ paying for good food), humans depend on past experiences and the opinions of their companions. As sentiment analysis uses data mining, it encapsulates public sentiment very accurately. Thus it is also the root of natural language processing, text analysis, and machine learning methods [1]. However, the shared data can be unstructured and have emotions or biases attached to it. Subsequently, many merchants are now moving towards sentiment analysis as it helps them promote their business and forecast a more accurate financial share of the market.

To understand the reviews and attitude of the reviews of movies, products, and restaurants, NLP is needed. There are several tool kits that support NLP tasks like NLTK, OpenNLP and Stanford CoreNLP. (Sun et al 2016). In NLP, tasks like tokenizing, stemming, padding, removing stopwords, POS tagging (Parts of speech), and named entity recognition (to recognize named entities in text data) are done . Named entity recognition is helpful for information extraction and preprocessing. Various machine learning algorithms are used such as a support vector classifier (SVC), a Decision Tree algorithm which are used to solve regression and classification problems. Also, few other algorithms like random forest algorithm, logistic regression, Multinomial Naive Bayes, and K nearest neighbors are used. To have better quality results from every algorithm, using a voting classifier, which is an ensemble method, is proposed. Further, RNN(Recurrent Neural Network) and LSTM (Long Short Term Memory Networks) are utilized to help achieve more accurate predictions. BERT (Bidirectional Encoder Representations from Transformers), which was developed very recently and trained on English Wikipedia (2500 million words) and Book Corpus (800 million words) from different genres was used. It has significantly improved the NLP tasks by presenting state-of-the-art results even for complex problems.[2] The tokens were tokenized using 37,000 WordPiece tokens.

The aim is to identify the sentiment of the review by deciding the polarity where the reviews are collected from different websites such as IMDb, Yelp, and Amazon. This can help the business evaluate their service and foresee public thoughts towards their movie, food, or products.

2. LITERATURE SURVEY
Presently, multiple methods of NLP are used for sentiment analysis, including deep learning networks or dictionary-based approaches. However, the dictionary has to be expanded due to the expansion of opinionated data, and it is not feasible to do so. A comprehensive collection of words and their sentiments is not easy to collect.

Hasan et al. (2019) [1] proposed a system with three major parts; Data extraction from Twitter and pre-processing it with NLTK library and then calculating the sentiment, where they have used the TF - IDF model to find out the essential words from the tweet to predict the sentiment. The Pickle module is used to build a classifier module, and the Pickle does the object serialization. This classifier model now calculates the sentiment of the tweet. It overrides the BoW model as it gives the same importance to all the words and does not reserve any semantic information.

Kanakaraj et al. (2015) [3] Note that techniques like BoW, NN combined SVM have limitations that can not be used uniformly across other applications. They gather their data through Twitter API v1.1 in a Tweet Objet format. Pre-processing of this data is

done to remove repeated words, clean the data by removing URLs, and extract features that will be tagged. In order to gather the semantic similarities between words, synonyms of WordNet are used. Before gathering synonyms, stemming is done to reduce the size of the feature vector and to preserve the key terms. These preserved key terms are now subjected to classifier and ensemble methods to achieve the sentiment of the text.

Solangi et al. (2018) [4] reviewed all the techniques of NLP. They pointed out the issues that were withstanding in NLP and recommended modifying the NLP techniques to close the communication gap between opinion and the engineering of the model. They suggest a 3 level process that works at the document level, sentence level, and fine-grained level. The difficulty and complex issues arise at the fine-grained level, but it becomes imperative at one stage to be very articulate with sentiment analysis.

Hu et al. (2019) [5] implemented a combination of word2vec and Bi - LSTM models. They have implemented one-way semantic analysis and multi-label sentiment analysis in the forward and backward text. Firstly, they generate word vectors using Word2vec. Bi - LSTM or Bi-directional LSTM do the forward and reverse calculation to extract the semantics.

Zheng et al. (2019) [6] proposes a dictionary-based convolutional recurrent network model to perform the sentiment analysis. To get the word vectors of the text and generate a dictionary, an automatic skip-gram model in the Word2vec network is used. This is used as an input for CNN to extract the local features required to create the sentence representation. This sentence representation is done by LSTMs. These sentence representations are given as input to a classifier to predict the sentiment of the text. The results were compared with traditional machine learning methods to optimize the parameters of the model.

(Ceci et al. 2016) [7] Proposed a model which is based on ontology. The research combined domain ontology with NLP techniques to understand the sentiment. They used movies and digital cameras for testing purposes. (Guerrero et al. 2015) [8] and (Cambria 2016) [9] assert that the NLP techniques are used to automatically detect the sentiment's polarity from text documents, online blogs, sentences, and words.

(Hemalatha and Ramathamika 2019) Implemented a supervised learning technique for sentiment analysis of a Yelp dataset. They included various algorithms like Linear SVC, Bernoulli Naive Bayes, Multinomial Naive Bayes, Naive Bayes and logistic regression. In their comparison it was deduced that Naive Bayes was most accurate with an accuracy of 79.12% and Bernoulli Naive Bayes was the least accurate with 73.22% accuracy. The paper also comments on the non usability of the system due its low accuracy.

Nair et al. (2021) compared logistic regression, BERT and VADER to analyse tweets that included Covid - 19. VADER stands for Valence Aware Dictionary and Sentiment Reasoner. VADER makes use of lexicons to get the valence (strength of words). VADER takes the polarity of each word and normalizes that polarity and applies it to the statement. The results showed that BERT in their case was more accurate than logistic regression while VADER and BERT were very close with their findings. The authors were able to get an accuracy of 0.92 in case of BERT. VADER was able to get accuracy of 0.88 while logistic regression was able to get 0.83 accuracy.

3. METHODOLOGY/ APPROACH

Machine Learning has proven to be a very promising tool in various fields such as biomedical engineering, stock market trading, analyzing text with the help of NLP. It has helped researchers implement various machine learning or deep learning algorithms to effectively and efficiently solve real-world problems. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) in Deep Learning have seen many advancements in recent times in almost every domain. A CNN learns to recognize patterns across space. On the other hand, RNN learns to recognize patterns across time. Natural Language processing is extensively used for sentiment analysis to gain insights into customer behavior and predict customer behavioral patterns for business models. To perform sentiment analysis on text data using machine learning algorithms, a lot of preprocessing and feature engineering is required. This can be overcome by using Deep Learning algorithms. An attempt was made to compare the performance of these two algorithms.

Natural Language Processing converts words (text data) into numbers. This is done as AI/ML algorithms understand numbers. These numbers are used to train the model to make predictions. The proposed sentiment analysis model can be used to understand public sentiment using public tweets, which will prove to be a high impact factor when deciding about the respective movie, commodity, or restaurant.

3.1. Dataset

The Sentiment Labelled Sentences Dataset is a publicly available dataset consisting of 3000 instances. Each instance is a sentence along with a score being a positive or negative sentiment. The dataset is collected from the UCI Machine Learning Repository archive. The score is set as either 0 for negative or 1 for positive. These sentences are collected from three websites/fields, namely IMDb, Amazon, and Yelp. For each website, there are 500 positive and 500 negative sentences. These sentences are picked from larger respective datasets with an attempt to select only positive and negative sentiments so that no neutral sentiments are collected. The instances used are text sentences extracted from reviews of movies, e-commerce products, and restaurants.

3.2. Preprocess the data

Preprocessing data is a crucial step in designing any model as the quality of the data affects the model's capabilities. In this process, using Label Encoder from the sci-kit learn library, the class labels can be converted to binary values. However, the class labels were already in the form of binary values. Data like some numbers or URLs and other symbols are taken care of using regular expressions and replaced with appropriate placeholders. Subsequently, in the data cleaning process, punctuations and whitespaces were discarded. Stop words are frequently used in English words like “this”, “that”, etc. These words do not contribute towards the classification task and can thus be discarded. Stopwords from the NLTK corpus library are utilized to remove the stopwords from the text data. To remove functional words such as pronouns and conjunctions, Parts of Speech (PoS) tagging is used. Gensim library is used for natural language processing and unsupervised topic modeling to preprocess the data. Similarly, using NLTK's Porter Stemmer, the word stems were removed.

3.3. Feature Engineering

This process had to be done for creating features for Machine Learning algorithms. In the data at hand, the words in each sentence are the features. To achieve this, each word has to be tokenized. The most common 1500 words or a bag of words can be used for features. A tokenizer can vectorize the text corpus. The way tokenization works is, each sentence is transformed into a sequence of integers. Padding is also used so that input

sequences are of the same length due to the constraints of Keras and TensorFlow as it accepts only fixed valued inputs. After this, the data is ready, divide it into the train and test sets. The dataset is divided into 80% train set and 20% test set.

4. APPLYING MACHINE LEARNING AND DEEPLEARNING TO PROCESSED DATA

4.1. Applying Machine Learning Algorithms to the Cleaned Data

For this task at hand, six different machine learning algorithms for the classification process of the text data.Logistic Regression, Support Vector Classifier, K Nearest Neighbors, Decision Tree classifier, Random Forest classifier, and Multinomial Naive Bayes classifier are used to make predictions. Logistic Regression is a mathematical model which helps to estimate the probability of occurrence of a feature. It works with binary data as it works with classification tasks. Multinomial Naive Bayes have a multinomial distribution that forms a feature vector which represents the frequency of occurrence of that specific feature. Decision trees can be used to categorize the given text to particular categories by running through the query structure from root to a certain leaf node which represents the document category. Random forest technique is an ensemble of decision tree learning method used for classification. It is an ensemble technique used to further improve the accuracy of the Bagged decision tree model. Random Forest classifier builds multiple decision trees models using the randomly selected subset of features of the training data [13]. Text classification problems generally have high dimension input space and are linear separable in nature. SVM's are largely used in automated text classification problem as they work well for high dimensional feature space [14]. On comparing the performance, the Support Vector classifier with a linear kernel outperforms the other algorithms. An ensemble method, i.e., the voting classifier, was used to extract the best features of all the six algorithms to get higher accuracy. Hard voting was used for the same.

4.2. Using Deep Learning on the Cleaned Data.

Feedforward neural networks or vanilla networks map a fixed size input- e.g., an image to a fixed-sized output- e.g., probabilities like in Regression or output classes in classification. This method does not have any memory effect or time dependency. A recurrent neural network is an artificial neural network that takes temporal dimension into consideration by using a feedback loop or some memory / internal state. A feedback loop means that, unlike vanilla neural networks, a temporal loop is present in which the hidden layer produces an output and feeds itself back. The temporal dimension mentioned is nothing but time. This enables the RNN to recall everything that happened in the previous timestamp, so the text sequence does not create any issues. The bi-directional LSTM model uses past and future data along with current information for a specific time frame. Long Short Term Memory Networks (LSTMs) have gates that can allow or block information from passing by. LSTMs are used rather than just RNNs because RNN suffers from vanishing gradient problems.

Similarly, bi-directional LSTMs are used rather than just LSTMs so that the model learns across the entire input rather than just previous time steps. The gates of LSTMs consist of a sigmoid neural network layer along with the pointwise application. Sigmoid activation is a binary function that either allows all data to flow or none of the data to flow.

4.3. Embedding Layer

The sequential model will have an embedding layer, bi-directional RNN and LSTM, and dense layers. The embedding layer learns a low-dimensional continuous representation of discrete input variables. In this, a number of low dimensional features that are relevant to represent the input data to avoid using a lot of resources to train the model can be mentioned. So basically, the layer learns to represent the original number of features (which can be huge) with fewer features that are used. This works similar to an autoencoder or Principal Component Analysis (PCA). This helps in effective learning by the layers. ReLU activation function is being used in the dense layers as it performs better than the hyperbolic tan function or traditional sigmoid function. Dropout was also used for better generalization of the model. Finally, the last dense layer has two classes, and softmax activation is used for the classification. The probabilities created at the end of this model are compared to the original labels using categorical cross-entropy.

5. BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a language model trained bidirectionally. It has proven to have a better and more profound sense of the context of language and flow compared to single-direction models. BERT was trained for 1M steps with a batch size of 128,000 words. BERT chooses a task-specific fine-tuning learning rate that performs the best on the development set. A transformer is used, which learns contextual relations between words in text data. The encoder reads the text input, and the decoder is responsible for prediction. Only the encoder is a part of it. The transformer reads the entire sequence of words at once, whereas directional models read words sequentially. This quality allows the model to understand word context based on the right and left of the word, i.e., based on its surroundings.

The input is a sequence of tokens from the dataset. They are first embedded into vectors then processed in the neural network. A sequence of vectors is obtained in the output. For the classification task, it is done in a way similar to Next Sentence Classification. It is done by adding a classification layer on top of the transformer output. Horev [2] explains, BERT's bidirectional approach (Masked Language Model) converges slower than left-to-right approaches (because only 15% of words are predicted in each batch). However, bidirectional training still outperforms left-to-right training after a small number of pre-training steps.

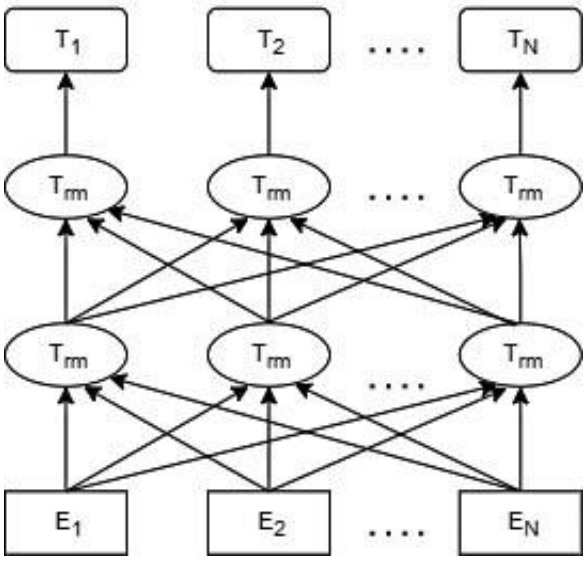

Figure 2: Bidirectional Encoder Representations From Transformers

5.1. Transformer

Attention mechanisms are applied here, which helps to gather information about the context of text data. Then the transformer encodes that context in a rich vector that represents the word smartly. Multi-head attention means that it computes 'h' different times with different weight matrices and then concatenates the results together.

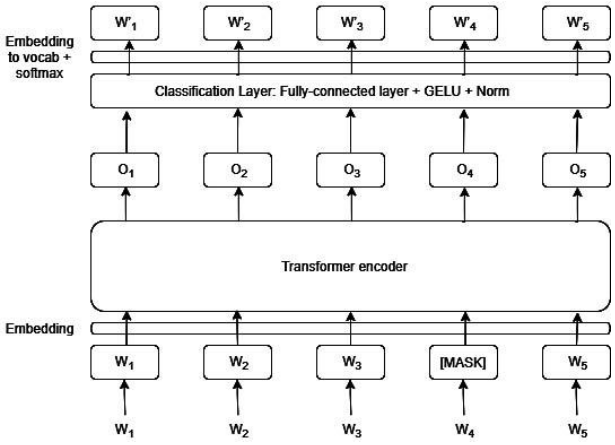

Figure 3: BERT Architecture

6. RESULTS

The dataset was tested on Machine Learning and Deep Learning algorithms. The implementation thus established a comparison between various algorithms that can be used to perform sentiment analysis. It is observed that the results are similar in some ways. Thus results can be improved using hyperparameter tuning in the used algorithms. Higher accuracy was achieved using the voting classifier than independent machine learning algorithms. The results of LSTMs were better than traditional machine learning algorithms. However, BERT was used due to the drawbacks of RNNs like vanishing gradient problem and single direction processing. Instead of pre-training the model on a language model, the BERT model was divided into two tasks: the "masked language model" and "next sentence prediction". In Next Sentence Prediction, when the model receives two sentences, its task is to predict if the second sentence follows the first in a corpus or not. The model was pre-trained in Next Sentence Prediction as the model could relate between two different sentences to perform tasks like natural language inference, and the Masked Language Model did not capture this knowledge. It is proved that pre-training with this second task notably increases performance in question answering and natural language inference.

Table 1: Comparison of results using different algorithms

Learning Method	Accuracy
Support Vector Machine	96.12%
Decision Tree	95.10%
Random Forest	95.20%
K Nearest Neighbours	94.10%
Multinomial Naive Bayes	95.80%
Logistic Regression	94.92%
Voting Classifier	96.23%
Long Short Term Memory Networks (LSTM)	96.70%
Bidirectional Encoder Representations from Transformers (BERT)	97.10%

7. CONCLUSION

In this paper, the model comprises different algorithms to get high accuracy consistently. With the help of traditional machine learning algorithms, the model was able to achieve around 95 percent accuracy, and higher accuracy was achieved with LSTM and even higher with BERT due to its advantages over RNNs and achieved an accuracy of 97.10 percent. Extensive work has been done in training a model using machine learning algorithms to compare the various accuracies. This output was further improved by the machine learning model using a voting classifier, which helps demonstrate each algorithm's effectiveness. Further, it is observed how LSTM and BERT have made advancements in NLP to increase the accuracy of the prediction of the sentiment of the text. The state-of-the-art BERT pre-trained model helped achieve an efficient model. It is observed that as BERT is pre-trained, its weights are learned in advance through the two unsupervised tasks: masked language modeling (predicting a missing word given the left and right context) and next sentence prediction (predicting whether one sentence follows another). Thus, BERT need not be trained from scratch when performing a new task. Instead, its weights are fine-tuned. For the classification purpose, the design is similar to the Next Sentence Prediction task. The model can be tested on larger datasets to generalize the model better. Different types of sentences can be used to diversify the data. Convolutional Neural Networks are also proving to be quite promising in the field of sentiment analysis. These algorithms can be tested in future studies to improve the accuracy and overall performance of the model.

8. REFERENCES

[1] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

[2] https://towardsdatascience.com/bert-explained-state-of-

the-art-language-model-for-nlp-f8b21a9b6270 (Accessed on 28 Aug 2021)

[3] M. Kanakaraj and R. M. R. Guddeti, "Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), 2015, pp. 169-170, doi: 10.1109/ICOSC.2015.7050801.

[4] Y. A. Solangi, Z. A. Solangi, S. Aarain, A. Abro, G. A. Mallah and A. Shah, "Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis," 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), 2018, pp. 1-4, doi: 10.1109/ICETAS.2018.8629198.

[5] J. Hu, X. Kang, S. Nishide and F. Ren, "Text multi-label sentiment analysis based on Bi-LSTM," 2019 IEEE 6th International Conference on Cloud Computing and Intelligence Systems (CCIS), 2019, pp. 16-20, doi: 10.1109/CCIS48116.2019.9073727.

[6] J. Zheng and L. Zheng, "A Dictionary-Based Convolutional Recurrent Neural Network Model for Sentiment Analysis," 2019 International Conference on Communications, Information System and Computer Engineering (CISCE), 2019, pp. 606-611, doi: 10.1109/CISCE.2019.00142.

[7] F. Ceci, A. L. Gonçalves and R. Weber, "A model for sentiment analysis based on ontology and cases," in IEEE Latin America Transactions, vol. 14, no. 11, pp. 4560-4566, Nov. 2016, doi: 10.1109/TLA.2016.7795829.

[8] Serrano-Guerrero, Jesus & Olivas, José & Romero, Francisco & Herrera-Viedma, Enrique. (2015). Sentiment analysis: A review and comparative analysis of web services. Information Sciences. 311. , 18–38. 10.1016/j.ins.2015.03.040.

[9] E. Cambria, "Affective Computing and Sentiment Analysis," in IEEE Intelligent Systems, vol. 31, no. 2, pp. 102-107, Mar.-Apr. 2016, doi: 10.1109/MIS.2016.31

[10] Shiliang Sun, Chen Luo, Junyu Chen,A review of natural language processing techniques for opinion mining systems,InformationFusion,Volume 36,2017,Pages 10-25,ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2016.10.004.

[11] H. S. and R. Ramathmika, "Sentiment Analysis of Yelp Reviews by Machine Learning," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 700-704, doi: 10.1109/ICCS45141.2019.9065812.

[12] A. J. Nair, V. G and A. Vinayak, "Comparative study of Twitter Sentiment On COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778, doi: 10.1109/ICCMC51019.2021.9418320.

[13] Xu, Baojun& Huang, Joshua & Williams, Graham & Wang, Qiang& Ye, Yunming. Classifying Very High-Dimensional Data with Random Forests Built from Small Subspaces. International Journal of Data Warehousing and Mining, 8(2), 44-63, 2012.

[14] Joachims, Thorsten. "Text categorization with support vector machines: learning with many relevant features." Paper presented at the meeting of the Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 1998.