APPROACHES OF CLASSIFICATION MODELS FOR SENTIMENT ANALYSIS

John Tsiligaridis

Department of Math and Computer Science, Heritage University,
Toppenish, WA, USA

ABSTRACT

Sentiment analysis (SA) is a Natural Language Processing (NLP*) method that helps identify the emotions in text. It is the automated process of identifying and classifying emotions in a text as positive, negative, or neutral sentiment. This way, companies can understand customers' sentiments, improve their products and services accordingly, and determine effective strategies. The need to discover the algorithm with the best classification performance is obvious. To this end, two different approaches for Sentiment Analysis problems are presented. The first one is based on Machine Learning (ML) models and the second one on Deep Learning (DP) models. Most ML models are flexible depending on their classifier hyperparameters and provide competitive accuracy levels but not all of them. Logistic Regression (LR), Random Forests (RF) of ML and the various models based on Neural Networks (NNs) of DL are applied. Useful results are obtained. Measures for classifiers' effectiveness are also provided.*

KEYWORDS

Random Forest, Machine Learning, Deep Learning

1. INTRODUCTION

A set of methods from ML and DL are used for IMDB classification. The LR with early stopping has the ability to break early in the training. RF can create different trees with different sub-features from the features. Better accuracy can be achieved using various hyperparameters. The CNN with the convolutional layer works as a filter that tend to excel at learning short sequences. For longest sequences, LSTM with its variances has also been used. ROC curves are used for performance estimation of the various classification approaches. Tuning hyperparameters were used for all the models. For example, CNN architecture accounted the number of filters, kernel length, dense layer architecture, dropout, number of hidden layers, while Regression classifier considered learning rate, epochs, etc.

The Logistic Regression (LR) method tries to estimate the probability p that the dependent variable will have a given value [1],[2],[3]. One technique to improve classification accuracy is by using Ensemble methods. Bagging, boosting, and random forests are popular ensemble methods [3]. Random Forest (RF) using bagging is an example of Ensemble Methods with the Decision trees (DT) as base classifier [4]. An ensemble combines a series of *k* learned models (or base classifiers), M_1, M_2… M_k, with the aim of creating an improved composite classification model, M [2]. An ensemble tends to be more accurate than its base classifiers. Ensembles yield better results when there is significant diversity among the models [2].

David C. Wyld et al. (Eds): AdNLP, CDKP, NCO, SAI, MLT, Signal, SOFT, ICAITA, CMC - 2024
pp. 81-86, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.141007

An ensemble for classification is a composite model, made up of a combination of classifiers. The ensemble based on the collection of votes returns the individual classifiers vote and a class label prediction. ROC curves are used for classification effectiveness. A ROC curve for a given model shows the trade-off between the true positive rate (TPR) and the false positive rate (FPR) [2]. AdaBoost is a popular boosting algorithm. An AdaBoost ensemble method with reduced entropy for Breast Cancer prediction is developed in [5]. In [6], a new ensemble ML method based on AdaBoost is developed for placement data classification analysis. Modern practical Deep Learning Networks with optimization and regularization are presented in [7].A Long Term Short Memory (LSTM) is presented, in [8], based Recurrent Neural Network (RNN), a popular deep learning algorithm for sentiment analysis of English and Spanish data. In [9] a deep (Long Short Term Memory) LSTM neural network for feature free classification of seven daily activities by using raw data collected from three-dimensional accelerometer. The CNN is described in [10] with introduction, overview, building blocks of CNN, different architecture of CNN, applications in several Domain areas, issues, and challenges. A novel deep learning model is proposed in [11], that combines multiple pipelines of convolutional neural network and bi-directional long-short term memory units for a stock price market application.

The paper organization is as follows. In Section 2 the ML models are included. Section 3 deals with Ensemble models. Section 4 contains the DL models. Simulation results are provided in Section 5.

2. ML MODELS

From the ML, the Logistic Regression (LR) algorithm takes the Bag of Words as input text in the form of a sentence and turns it into a feature vector. A sentiment classifier using LR is built with the goal of having a linear function with all the dependent variables and their associated weights as the parameter of the sigmoid function. The training phase starts with learning rate of 0.1. Early stopping and the sigmoid function are also applied. The early stopping take place during training when the error rate changes to values lower than a predefined threshold.

The effectiveness of LR classifier is achieved by using a ROC curve. The ROC curve is a tool commonly used with binary classifiers. It plots TPR (true positive rate) against FPR (false positive rate). The ROC curve for the logistic regression sentiment classifier is 87%. In addition, the SoftMax regression (SMLR) for more than two classes and for generalization purposes but provides no better accuracy than LR. The softmax function is applied instead of the sigmoid to learn the probability distribution over the N classes we are trying to predict. LR belongs to the group of inflexible algorithms since it cannot adapt or learn from data.

By switching to the nonlinear algorithm, the Random Forest (RF), a higher level of accuracy can be achieved having just to increase the number of estimators. It builds a collection of Decision Trees(DTs) in order to mitigate the risk of uncertainty. Random Forest(RF) builds multiple decision trees and merges them together to get a more accurate and stable prediction. It reduces the overfitting of datasets and increases precision of the outcome. The bagging aggregation (works with replacement) is used for the training dataset and it creates different samples of data. The outcome chosen by most DTs will be the final choice(majority vote).With RF, the selection of tuning parameters for optimization (i.e. number of trees etc.) is made using grid search algorithm.

3. ENSEMBLE MODELS

A new dimension for the ML models is provided by developing ensemble models. Ensemble model is a ML algorithm that aggregates the predictions of multiple estimators or models. The purpose of an ensemble model is to provide better predictive performance than any single contributing model. Tuning parameter methods are applied before proceeding with the ensemble models.

To increase the performance of RF a sequential ensemble model is used. Sequential ensemble methods such as busting aim to combine several weak learners into a single strong learner.

The Random Forest is an ensemble model using bagging as the ensemble method and use the Decision Tree (DT) as the individual model. The Tree Based ensemble models considering the DT as base model is the most frequently used.

AdaBoost is a boosting ensemble model that cooperates with DTs. It learns from the mistakes by increasing the weight of misclassified data points [12]. Adding a boost to RF with AdaBoost (strong learner) better performance is achieved. The accuracy with AdaBoost became: 88% compared to the initial RF after selecting the appropriate tuning parameters accuracy of 83.5%.

4. DL MODELS

A group of Deep Learning (DL) algorithms are used for classification starting from a Simple Dense Neural Network with Embedding layer (DNE) and continuing with the Convolutional Neural Network (CNN), the Long-Short Term Memory (LSTM), the LSTM with Dropdown (LSTMD), and the Ensemble model with LSTM and CNN(LSTM_CNN).

Hyperparameters tuning for the NNs (i.e. learning rate, etc.) are also used. Finally, the grid search method is used for discovering the LSTM hyperparameters for better accuracy. CNN architecture considered the number of filters, kernel length, dense layer architecture (=256), dropout (20%), number of hidden layers.

The DNE is the simplest Neural Network (NN) model to classify film reviews by their sentiment. The DNE architecture includes among others: the sequential method, and the Embedding layer that enables the creation of word vectors from a corpus of documents (25,000 movie reviews of the IMDB training dataset).The problem of the DNE classifier is that it is not so suitable for detecting patterns of multiple tokens that predict the sentiment of the film review. There is misclassification of false positive and false negative.

To improve the performance of DNE the CNN is used. CNN is used to detect spatial patterns among words. For this purpose, a one-dimensional convolutional layer is provided along with the appropriate hyperparameters for the convolutional layer architecture (i.e. filters). The relu activation function is used in the convolutional layer. This layer has filters whose purpose is to activate when it meets a particular three token sequence.

Short sentiment sequences such as triplets of words can be identified, and this way learning can be improved. For a natural language document such as the movie review might contain much longer sequence of words that if considered altogether could predict some more outcomes.

To manage long sequences of data the LSTM that comprises specialized layers of the long short memory units can be used. LSTM receives input from a sequence of data, and it also receives input from a previous time point in the sequence.

There are two linear transformations that show where a cell in an LSTM layer can add information to the cell state which will be passed onto the next cell in the layer. The sigmoid activation with values of 0 or 1 operates as "gates" in order to decide if a new information should be added to the cell state. LSTM provided not so good performance despite its relative sophistication.

To avoid overfitting the LSTM dropout (LSTMD) is proposed. Dropout can be applied between layers using Dropout Keras layer. For this purpose, Dropout layers between the Embedding and LSTM layers and the LSTM and Dense output layers are added. The CNN is used for learning spatial structures and it can pick out invariant features for the two types of sentiment. This learned spatial features can be learned with a LSTM layer.

The LSTM_CNN model can be considered from the embedding layer followed by one dimensional CNN with the max pooling layers and finally the LSTM.

5. SIMULATION

The simulation is based on various experiments.

1.Accuracy for DL algorithms: DNE, CNN, LSTM, LSTMD, LSTM_CNN.
CNN has the best performance (Figure 1)

2. Accuracy for ML algorithms: LR, RF, RF_ADA (Figure 2)
RF_ADA has superiority over ML algorithms, and it is just the second in accuracy after the CNN.

3. ROC curve for the logistic regression sentiment classifier. The ROC/AUC is 87% (Figure 3)

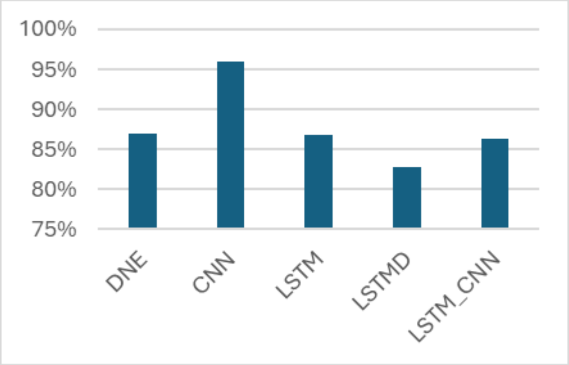

Figure 1. Accuracy for DL models.

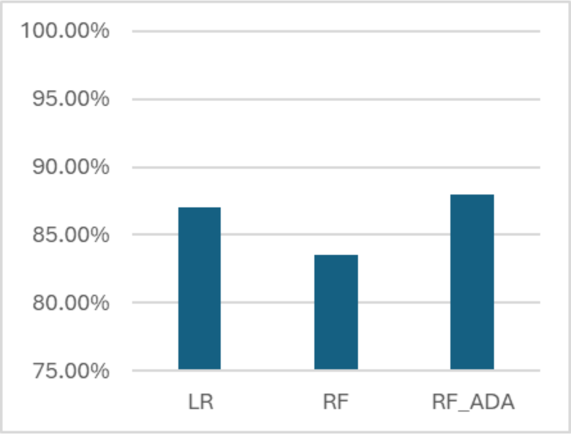

Figure 2. Accuracy for ML models

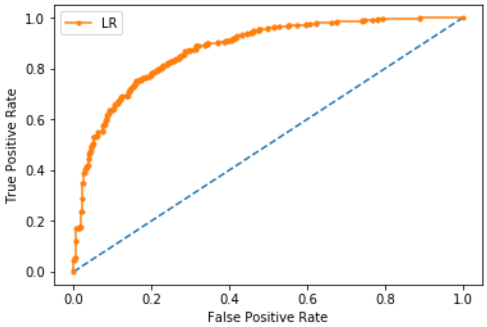

Figure 3. ROC curve for logistic regression

6. CONCLUSION

Sentiment Analysis (SA) for NLP methods have been developed. For the ML approach, accuracy has been achieved with the LR along with the SoftMax regression also providing ROC curves. The accuracy of LR provided better performance compared to SoftMax regression. The RF_ADA, the ensemble model, outperforms LR. RF_ADA surpasses all the proposed models except for CNN. CNN with the convolutional layers provided better accuracy than the DNE and the other DL algorithms.

From the comparison of the two sets of models the selection of the models with their hyperparameters for creating an ensemble model could be an issue if other algorithms with their tuning parameters can provide good performance.

The sequential Ensemble method (AdaBoost) RF_ADA provides superior results than the base model Decision Trees (DTs) and the simple RF.The use of Ensemble models using any ML model as base estimator (i.e. DTs) can improve performance and become competitive to some DL models after selecting the appropriate hyperparameters.

Generally, there is a superiority of DL models over ML models because of their flexibility since there are many different types with various kinds of hyperparameters (i.e. hidden layers, convolutional layers, dropout).

REFERENCES

[1] Karntardzic, M. (2020). *Data Mining: Concepts, Models, Methods, and Algorithms*, IEEE Press.

[2] Han, J., Kamber, M., Pei, J. (2012). *Data Mining Concepts and Techniques,* Morgan Kaufman, 3rd edition.

[3] Gareth, J., Witten, D., Hastie, T., Tibshirani, R. (2017*). An Introduction to Statistical Learning with Applications in R*, Springer, 2nd edition.

[4] Rokach, L. (2009). *Ensemble Learning: Patern Classification Using Ensemble Methods,* World Scientific, 2nd edition.

[5] Ramakrisna, M., Venkatesan, V., Izonin, I., Havryliuk, M., Bhat, C.(2023). Homogenous Adaboost Ensemble Machine Learning Algorithms with Reduced Entropy on Balanced Data. *In Entropy. Basel, Switzerland*,*25*(2),245. https://doi.org/10.3390/e25020245

[6] Kalaiselvi, B., Geetha, S. (2022). Ensemble Machine Learning AdaBoost with NBtree for Placement Data Analysis. *In 2nd International Conference on Intelligent Technology (CONIT).* Hubli, India. doi:10.1109/CONIT55038.2022.9847993

[7] Goodfellow, I. Bengio, Y., Courville, A. (2016). *DeepLearning,*The MIT Press.

[8] Saha, B., Senapati, A. (2020). Long Short-Term Memory (LSTM) based Deep Learning for Sentiment Analysis of English and Spanish Data. *In International. Conference on Computational Performance Evaluation (ComPE).* Shillong, India.

[9] Guney, S., Erdas, C. (2019). A Deep LSTM Approach for Activity Recognition. *In 42nd Intern. Conference on Telecommunications and Signal Processing (TSP),* Budapest, Hungary.

[10] Pandey, K., Patel, S. (2023). Deep Learning with Convolutional Neural Networks: from Theory to Practice. *In 7th International Conference on Trends in Electronics and Informatics (ICOEI)*Tirunelveli, India.

[11] Eapen, J., Bein, D., Verma A. (2019).Novel Deep Learning Model with CNN and Bi-Directional LSTM for Improved Stock Market Index Prediction. *In IEEE 9th Annual Computing and Communication Workshop and Conference(CCWC*), Las Vegas, NV, USA.

[12] Chan,J., Jayasuriya,D., Sundaram,D. (2020). *Machine Learning Applications in Healthcare: Breast Cancer Diagnosis and Prognosis,* SSRN, Elsevier.http://dx.doi.org/10.2139/ssrn.4211998.