Performance Analysis of Machine Learning Algorithms for Movie Review

Arafat Habib Quraishi
Department of Computer Science and Engineering
Leading University, Sylhet-3112, Bangladesh

ABSTRACT
In this paper, we have evaluated the performance of four machine learning algorithms in terms of sentiment analysis in the IMDB review dataset. Among these algorithms, two are neural network based and two are non-neural network based. We used binary classification for sentiment analysis in IMDB reviews and examined all the four algorithms to detect whether the sentiment of the text is positive or negative. Among the neural network based approaches, we applied LSTM and GRU. We found that GRU performed better than LSTM. Among the non-neural network based algorithms, we applied Multinomial Naïve Bayes and Support Vector Machine. We found that SVM outperformed Multinomial Naïve Bayes. Among these four algorithms, GRU performed the best with an accuracy of 89.0%.

Keywords
IMDB, Reviews, Sentiment Analysis, Neural Network, LSTM, GRU, SVM, Naïve Bayes.

1. INTRODUCTION
With the rise of technology, user can directly give review about products, brands etc. These reviews play vital role in online shopping as well as help people to determine whether a product is good or not. These reviews or opinions play significant impact on the success of a business. It helps companies to improve their products based on analyzing the user feedback. Thus it is very important to correctly analyze this huge amount of text in an effective and accurate way.

Huge amount of text data related to user opinions about products and services are generated every day in the world. But it is not feasible to analyze the sentiment of these vast of texts manually. So, an automated process must be applied to mine these text data and analyze the sentiment effectively as the companies need to use these numerous amounts of data to improve their businesses by drawing more effective marketing analysis, product reviews, public relations etc.

Sentiment analysis is the process of computationally identifying sentiments expressed in a text, especially in order to determine whether the writer's attitude towards a particular topic or product is positive or negative or neutral. Different natural language processing or text analysis techniques are applied for sentiment analysis.

There are many publicly available datasets such as Amazon Reviews, IMDB Movie Reviews, Yelp Reviews, etc. which are used by researchers to investigate sentiment analysis techniques. Previously various feature engineering-based approaches have been applied for sentiment analysis. But these approaches require handcrafted features which are mostly error prone. With the rise of Machine Learning and Deep Learning, various models provide new state-of-the-art results.

In this paper, we are motivated to use four Machine Learning and Deep Learning based models for sentiment analysis to evaluate their performances in the IMDB review dataset. There are four algorithms we have applied for sentiment analysis. Two of these algorithms are neural network based: Long Short-Term Memory Model (LSTM) and Gated Recurrent Unit (GRU) and the other two are non-neural network based: Multinomial Naïve Bayes and Support Vector Machine. We have done Binary Classification in the IMDB movie review dataset to evaluate their performances.

2. RELATED WORK
A significant number of researches have been conducted in the field of sentiment analysis in recent years.

Earlier, various rule-based approaches have been used for sentiment analysis. For example, Hutto and Gilbert [4] presented a simple rule-based model for general sentiment analysis and found better performance than the benchmarks used in their study. But the performance of their proposed model was not compared with neural network based approaches. Popular Social Media website like Twitter has also been used for sentiment analysis. Agarwal et al. [1] examined sentiment analysis on Twitter data by introducing Parts of Speech (POS) features. Later, Kouloumpis et al. [5] investigated the utility of linguistic features for detecting the sentiment of Twitter messages and showed that part-of-speech features is not useful for sentiment analysis in the micro blogging domain. Wilson et al. [15] presented a new approach to phrase-level sentiment analysis that first determined whether an expression was neutral or polar, and then disambiguated the polarity of the polar expressions.

Different techniques in combination with sentiment analysis algorithms have also been applied. Liu et al. [7] applied sentiment analysis models for predicting the helpfulness of reviews, which provides the basis for discovering the most helpful reviews for given products. Reviewers review history was also considered by some researchers. For example, Basiri et al. [2] considered the comment histories of reviewers and found that their proposed system performed better than different algorithms. But they only compared their model with Machine Learning based algorithms and did not compare the performance of their proposed model with neural network-based approaches.

Various unsupervised approaches were proposed for sentiment analysis. Lin and He [6] proposed an unsupervised probabilistic modeling framework based on Latent Dirichlet Allocation (LDA), called joint sentiment/topic model (JST), which could detect sentiment and topic simultaneously from text. Turney [13] introduced an unsupervised learning algorithm for sentiment classification based on semantic orientation of the phrases using the PMI-IR score. His presented approach performed well in reviews from different domains but for movie reviews the performance was bad. Turney and Littman [14] introduced a method for inferring the semantic orientation of a word from its statistical association.

Among supervised models, Melville et al. [9] developed an

effective framework for incorporating lexical knowledge and successfully applied the developed approach to the task of sentiment classification. Paltoglou and Thelwall [10] found that variants of the classic TF-IDF scheme adapted to sentiment analysis provided significant increases in accuracy.

Maas et al. [8] presented a model that used a mix of supervised and unsupervised techniques to learn word vectors capturing semantic term and sentiment content. Their proposed model outperformed many previously introduced sentiment classification techniques.

Various machine learning algorithms were also proposed for sentiment analysis. Pang et al. [11] employed three different machine learning methods (Naïve Bayes, Maximum Entropy Classification, and Support Vector Machines) for document level sentiment analysis and found that their techniques outperformed human-produced baselines. In medical domain, Laskar et al. [16] and Yadav et al. [17] applied various machine learning based approaches such as SVM, Decision Tree, and etc. and showed good performance.

Prabowo and Thelwall [12] proposed a hybrid approach for sentiment analysis by combining rule-based classification with supervised learning and machine learning. They applied this hybrid approach in movie reviews, product reviews and MySpace comments and found that hybrid approach could improve the classification effectiveness.

Among Deep Learning based approaches, most notable work is by Devlin et al. [18] which showed very impressive performance in different Natural Language Understanding task including sentiment analysis. They used the encoder of the Transformer Model and showed state-of-the-art performance.

In this paper, we have applied both Deep Learning and Machine Learning based approaches for sentiment analysis and compared their performances.

3. METHODOLOGY

We applied four different algorithms for sentiment analysis in IMDB review dataset. Two of them are neural network based and others are non-neural network based. All the algorithms that we used are briefly described below.

3.1 Multinomial Naïve Bayes

Multinomial Naïve Bayes is a simple classification method based on Bayes rule. It relies on simple bag of words representation of documents or texts.

For a document or text *d*, and class *c*, Naïve Bayes predicts the probability of the class *c* for text *d* with the following conditional probability:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

After applying Naïve Bayes rule, we get:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x|c) \quad (2)$$

For Naïve Bayes Multinomial, we integrated the TF-IDF weighting to generate the list of words. We implemented the Multinomial Naïve Bayes in Python using the Scikit-learn library.

3.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised learning algorithm. SVM has shown great performance in classification tasks. Not only classifying linearly separable data, the SVM models are also able to efficiently handle a non-linear classification problem using the kernel function which transforms the low dimensional input data to relatively higher dimensional spaces. The models can efficiently handle high dimensional feature vectors. In this regard, the SVM has the great potential as a solution for the sentiment classification task.

3.3 Long Short-Term Memory Model

Long Short-Term Memory Model (LSTM) units are units of Recurrent Neural Network (RNN). LSTM networks are used for classification or prediction based on time series data. It can deal with exploding and vanishing gradient problems. One of the major problems of RNN is the long-term dependency [3]. LSTM can avoid such long-term dependencies. A common LSTM unit is composed of a cell, input gate, output gate and a forget gate. In Figure 1, a simple LSTM network is shown.

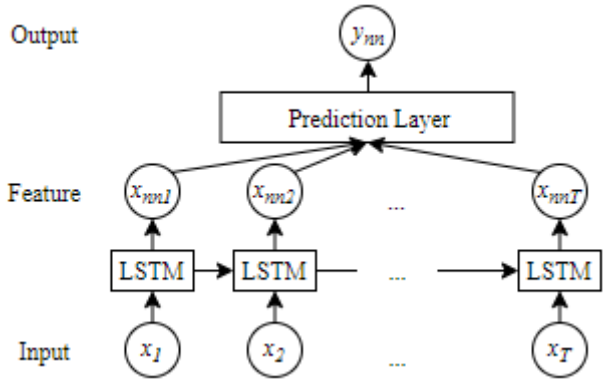

Figure 1. LSTM Network

For our experiment, we used one LSTM layer with 100 hidden nodes, a dropout rate of 0.2, 'adam' as the optimizer, the sigmoid function as the activation function and 'early stopping' during training. We implemented the LSTM architecture in Python using Keras library. We also used Keras embedding to generate word vector of for the embedding layer.

3.4 Gated Recurrent Unit

Gated Recurrent Unit (GRU) is a simplified version of Long Short-Term Memory (LSTM) model. Even though GRU has fewer parameters, the model is able to efficiently capture long term dependencies between sequences. Therefore, GRU is comparable to LSTM in terms of performance and computational efficiency [7]. In this regard, GRU model can be used as one of the solutions in the classification task. We implemented the deep neural architecture using Keras library. The architecture includes 3 layers: 1) Word embedding layer, 2) GRU layer, 3) Fully connected layer with activation function. Input data was 'Bag of Words' vectors from reviews. For activation function in fully connected layer, we used sigmoid function for binary classification. Dropout rate was 0.2.

4. DATASETS

We used the IMDB movie review dataset to see the performance of our algorithms for the task of sentiment analysis. In this dataset, there are 12500 positive and 12500 negative instances in the training set. In the evaluation set, there are also 12500 positive and 12500 negative instances.

The negative and positive examples in IMDB movie review dataset is shown on Figure 2 and Figure 3 respectively.

Text: *Yep, Edward G. gives us a retro view of the criminal defense world. First he's an overzealous prosecutor who sends the wrong man to the chair (played passionately, albeit briefly by DeForrest Kelly), then he's so filled with remorse his only solace is the bottle. Throw in a jaded romance, a genuinely rapid descent into penury and no qualms about who he defends, and next thing you know -- Shazam! Black Leg Lawyer (god I love that phrase). He sees the light just in time to save his jaded beloved from the chair. Yawn. But really, the courtroom action is pure melodrama. See him punch out a witness, see him drink poison, see him argue passionately as he clutches a bullet hole in his breast. Be prepared for melodrama. The hoot of the film though, is Jayne Russell. With curves defying the laws of gravity and an IQ approached absolute zero, she is something to see. Even sings a bit. 0*
Label: *0*

Figure 2. Negative example in IMDB movie review dataset

Text: *How many movies are there that you can think of when you see a movie like this? I can't count them but it sure seemed like the movie makers were trying to give me a hint. I was reminded so often of other movies, it became a big distraction. One of the borrowed memorable lines came from a movie from 2003 - Day After Tomorrow. One line by itself, is not so bad but this movie borrows so much from so many movies it becomes a bad risk. BUT... See The Movie! Despite its downfalls there is enough to make it interesting and maybe make it appear clever. While borrowing so much from other movies it never goes overboard. In fact, you'll probably find yourself battening down the hatches and riding the storm out. Why? ...Costner and Kutcher played their characters very well. I have never been a fan of Kutcher's and I nearly gave up on him in The Guardian, but he surfaced in good fashion. Costner carries the movie swimmingly with the best of Costner's ability. I don't think Mrs. Robinson had anything to do with his success. The supporting cast all around played their parts well. I had no problem with any of them in the end. But some of these characters were used too much. From here on out I can only nit-pick so I will save you the wear and tear. Enjoy the movie, the parts that work, work well enough to keep your head above water. Just don't expect a smooth ride. 7 of 10 but almost a 6.*
Label: *1*

Figure 3. Positive example in IMDB movie review dataset

5. TRAINING AND PARAMETER SETTINGS

For our neural network based approaches, we used GloVe [19] word embedding. Maximum 300 tokens were considered. We ran 3 maximum Epoch. We describe the results in the following sections.

5.1 Classification Results

Results of all the four models are given on Table 1.

Table 1. Classification Results

Model Name	Accuracy
Multinomial Naïve Bayes	83.5%
SVM	88.3%
LSTM	88.5%
GRU	89.0%

5.1.1 Multinomial Naïve Bayes

We evaluated the performance of Multinomial Naïve Bayes for binary classification. Multinomial Naïve Bayes performed the worst among non-neural network based approaches in terms of accuracy. It also showed the worst performance among all the four algorithms used in this paper, though the computation time of Naïve Bayes is very efficient. The accuracy is 83.5%

5.1.2 SVM

For binary classification, SVM showed the best performance among the two non-deep learning based models, and the third best performance among all the four models. The accuracy is 88.3%.

5.1.3 LSTM

LSTM performed better than all the non-neural network based approaches in terms of binary classification. The accuracy of LSTM for binary classification was very close to our best model. The accuracy is 88.5%.

5.1.4 GRU

For binary classification, GRU performed the best among all the four models. It gave slightly better performance than our second best model in terms of accuracy. The accuracy is 89.0%.

6. DISCUSSIONS

The overall result of our models in terms of binary classification is shown on Figure 4.

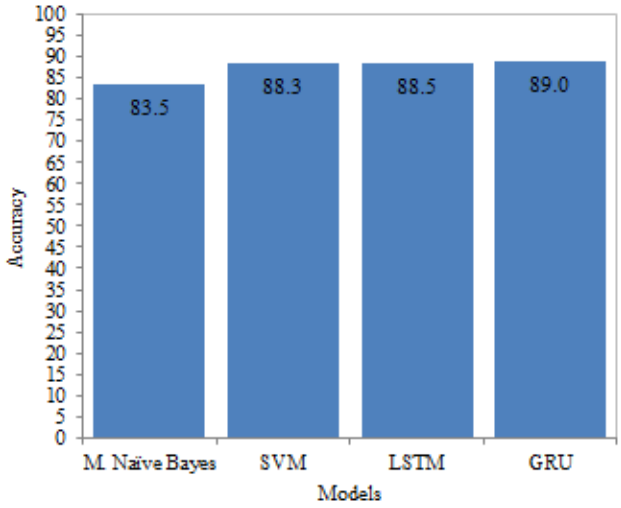

Figure 4. Accuracy (%) of each model in Binary Classification

From Figure 4, we can see that for binary classification, all the neural network based approaches provided more than 85% accuracy. GRU performs the best with 89.0% accuracy and Multinomial Naïve Bayes performs the worst with 83.5% accuracy. The accuracy of LSTM is 88.5%. Among the non-neural network based approaches, SVM performs the best with 88.3% accuracy.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we applied four different algorithms for sentiment analysis in the IMDB review dataset. We found that all the neural network based approaches outperformed the

non-neural network based approaches in terms of binary classification. We implemented all the neural network based algorithms in Python using Keras library. The Multinomial Naïve Bayes and SVM were also implemented in Python using Scikit-learn library. For binary classification, GRU performed the best with 89.0% accuracy. In future, these models can be evaluated on more datasets. Here, we did not apply these models for multiclass classification. So, the future research can be done on evaluating performances of these models in terms of accuracy for multiclass classification in other datasets.

8. ACKNOWLEDGEMENTS

Thanks to M. T. R. Laskar for providing guidance with helpful comments throughout the development of this project.

9. REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media, LSM '11, ACL, 30-38.

[2] Basiri, M., Ghasem-Aghae, N., & Naghsh-Nilchi, A. (2014). Exploiting reviewers' comment histories for sentiment analysis. Journal of Information Science, 40(3), 313-328.

[3] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 157-166.

[4] Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In International AAAI Conference on Weblogs and Social Media, AAAI, 216-225.

[5] Kouloumpis, E., Wilson, T. & Moore, J., (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! In Proceedings of the Fifth International Conference on Weblogs and Social Media, 538-541.

[6] Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, ACM, 375-384.

[7] Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and Predicting the Helpfulness of Online Reviews, in Eighth IEEE International Conference on Data Mining, Pisa, 443-452.

[8] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistic, ACL, 142-150.

[9] Melville, P., Gryc, W., & Lawrence R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09, ACM, 1275-1284.

[10] Paltoglou, G., & Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10. ACL, 1386-1395.

[11] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, EMNLP '02, ACL, 79-86.

[12] Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. Journal Of Informetrics, 3(2), 143-157.

[13] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, ACL, 417-424.

[14] Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems (TOIS), 21(4), 315-346.

[15] Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT '05, ACL, 347-354.

[16] Laskar, M. T. R., Hossain, M. T., Kamal, A. R. M., & Rashid, N. (2016). Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction. International Journal of Computer Applications, 975, 8887.

[17] Yadav, S., Ekbal, A., Saha, S., & Bhattacharyya, P. (2018, May). Medical sentiment analysis using social media: towards building a patient assisted system. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

[18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[19] Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).