A Comparative Study on Different Neural Networks on Sarcasm News Headlines Detection

Yuqing Qiao
Khoury College of Computer Sciences
Northeastern University
qiao.yuqi@northeastern.edu

Zhenyu Wang
Khoury College of Computer Sciences
Northeastern University
wang.zheny@northeastern.edu

Jingjing Lin
Khoury College of Computer Sciences
Northeastern University
lin.jingj@northeastern.edu

Abstract

In this study, we present a comparative analysis of sarcasm detection in news headlines using various machine learning models. Specifically, we reimplement the BERT (Bidirectional Encoder Representations from Transformers) model and compare its performance with traditional models such as Multilayer Perceptron (MLP), logistic regression, and Naive Bayes classifiers. Our experiments involve training and testing these models on a dataset of sarcasm-labeled news headlines. Results indicate that BERT outperformed the traditional models such as MLP, logistic regression, and Naive Bayes. Through this study, we aim to provide insights and gain a deep understanding of the encoder used in BERT. The source code used for this study is available in the repository [1] and the dataset used in our experiments can be found at [2].

1 Introduction

Sarcasm detection poses a significant challenge within the domain of Natural Language Processing (NLP). The objective of this project is to examine the efficacy of different machine learning models in the identification of sarcasm within text. Unlike straightforward content, sarcasm requires the interpretation of context, tone, and often, cultural or situational knowledge, making it a complex linguistic construct for algorithms to decipher.

This report details the implementation and evaluation of various models including Naive Bayes, Logistic Regression, Multilayer Perceptron (MLP), and the transformer-based Bidirectional Encoder Representations from Transformers (BERT). Each model's capacity to process and classify sarcastic statements within a dataset of news headlines is rigorously tested. The comparative study aims to not only quantify the success rates of these diverse models but also to understand their respective strengths and limitations in processing nuanced linguistic features.

2 Method

We reimplemented the Naive Bayes, Logistic Regression, Multilayer Perceptron (MLP), and the BERT architect as best as we could, we consulted other implementation code [3], medium explaination blogs and other open source references.

2.1 Dataset and Data Preprocessing

The dataset, sourced from the Sarcasm Headlines Dataset, was first loaded into a Pandas DataFrame. The data was cleansed of null values to maintain consistency and quality, following which the headlines were extracted as textual input for the models and the sarcasm labels were prepared as targets. A key part of the preprocessing involved splitting the dataset into training and validation sets to facilitate model training and subsequent performance evaluation.

2.2 Architecture

2.2.1 Text Vectorization

For traditional machine learning models like Naive Bayes, Logistic Regression, and MLP, the textual data was transformed into TF-IDF (Term Frequency-Inverse Document Frequency) feature vectors. The TextVectorization layer from TensorFlow was used to standardize and tokenize the text data. This step was critical for converting text into a numerical format that machine learning models could process and learn from.

2.2.2 NB, LR, MLP

A Multinomial Naive Bayes model was implemented using scikit-learn and trained on the TF-IDF vectors. The Logistic Regression model had the max_iter parameter set to 1000 to ensure convergence during training.The MLP used a specific architecture with two hidden layers sized 50 and 20 nodes respectively.

2.2.3 BERT

The BERT language model consists of two parts which are masked language model (MLM) and next sentence prediction (NSP). During training, random words in the text are masked, and the MLM is trained to utilize the context before and after each masked word, employing the attention mechanism to predict the masked words. NSP, on the other hand, is trained to predict the sequence order of sentences. Correct sentence pairs are mixed with random sentence pairs, enabling BERT to predict the correct next sentence.

After pretraining the BERT model, it is fine-tuned for downstream tasks. The fine tuning for binary classification utilized the pre-trained bert-base-uncased model provided by Hugging Face's transformers library. The training utilized a custom DataLoader to process the tokenized data in batches of size 16. The model was trained for 1 epoch with a learning rate of 2e-5 using the AdamW optimizer.

The BERT model's predictions were decoded using a softmax layer, which converted the logits to probabilities. This process took place during the evaluation where predictions on the validation set were made.

The BERT model utilized a batch size of 16 and was trained for 1 epoch, with a learning rate of 2e-5, using the AdamW optimizer. These parameters were chosen to balance the computational efficiency against the training effectiveness.

The training process involved utilizing the DataLoader to iterate over batches of tokenized input data and corresponding labels. For BERT, the input data included attention masks and input IDs essential for the transformer model's processing.

Each model was trained on the vectorized text data. The BERT model required additional preprocessing steps, utilizing the BertTokenizer to process the text into a format suitable for the BERT architecture. This included attention masks and input IDs, essential for the model to understand and weigh the input sequence's various components.

3 Results

The training process across models included several consistent steps, with the primary focus on loss reduction and accuracy improvement over epochs. Each model's performance was rigorously documented using a variety of metrics.

Accuracy is measured as the proportion of correctly predicted instances over the total number of instances.

Precision and Recall is evaluated for both sarcastic and non-sarcastic classes, providing insight into the models' ability to correctly identify and classify each class.

F1-Score is calculated as the harmonic mean of precision and recall, offering a single metric for the balance between these two metrics.

The evaluations for Naive Bayes, Logistic Regression, and MLP were carried out using scikit-learn's metric functions after the models predicted the validation set's classes.

BERT: the evaluation involved running the validation set through the fine-tuned model and collecting predictions. The performance metrics were computed and compared against those of the other models. BERT's training process was detailed, highlighting the use of a custom DataLoader to handle batches of tokenized text data, the training loop with loss computation and optimization steps, and a final evaluation to determine the model's accuracy and produce a detailed classification report.

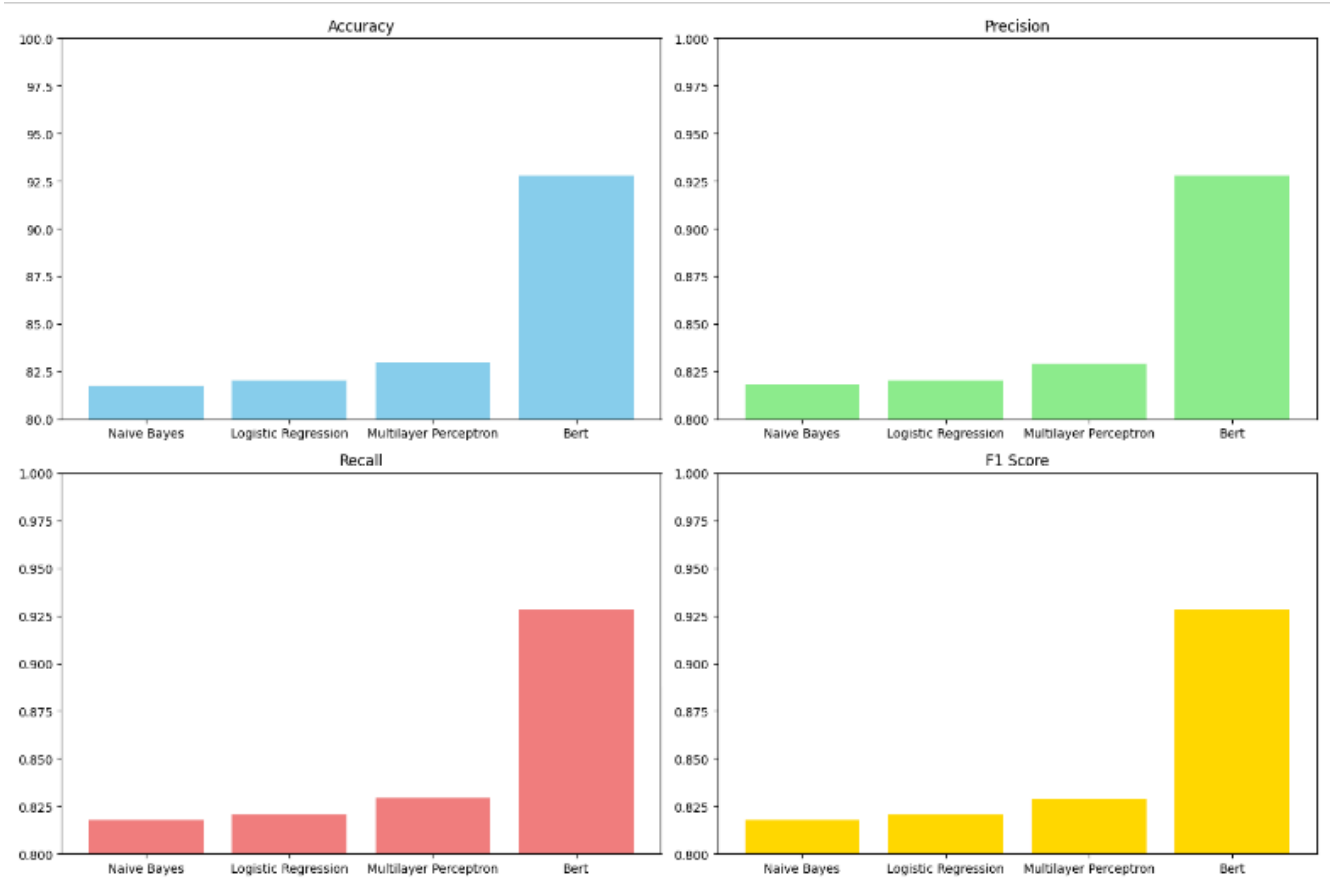

Figure 1: (a) Naive Bayes exhibited an accuracy of 81.77%, with a precision, recall, and F1 score of 0.818. This indicates a balanced performance but highlights limitations in handling complex language constructs like sarcasm. (b) Logistic Regression achieved slightly better accuracy at 82.05%, and the precision, recall, and F1 score mirrored the accuracy, suggesting consistent performance across various aspects of prediction. (c) Multilayer Perceptron (MLP) performed better than the previous two models with an accuracy of 82.95%, and equal precision, recall, and F1 score of 0.829, indicating its deeper network was able to capture more nuanced patterns in the data. (d) BERT significantly outperformed the traditional models with an impressive accuracy of 92.79%, and a weighted average precision, recall, and F1 score of 0.928. These results demonstrate the advanced capabilities of the BERT model in contextual understanding and sarcasm detection.

Post training, the evaluation on the validation set was conducted, yielding an accuracy of 92.79% for the BERT model. This suggests that BERT was highly effective in correctly classifying sarcastic and non-sarcastic headlines. The precision and recall metrics were notably high for the BERT model, with a weighted average precision of 0.928 and a weighted average recall of 0.928. This balance indicates the model's ability to correctly identify positive samples (precision) and its ability to find all positive samples (recall). The F1-score, a harmonic mean of precision and

recall, was also 0.928 for BERT, underscoring the model's robust performance across these metrics.

These detailed figures provide a clearer view of the BERT model's training regime and its subsequent validation performance, as documented in the provided PDF(code). The BERT model outperformed the traditional machine learning models on all accounts, which were also trained and evaluated in the same context using the scikit-learn library's implementations. The Naive Bayes model, Logistic Regression model, and MLP model all had their own set of hyperparameters and were evaluated using accuracy, precision, recall, and F1-score metrics, similar to the BERT model. However, their performance was lower, as they achieved accuracies of 81.77%, 82.05%, and 82.95% respectively, as compared to BERT's 92.79%.

4 Conclusion

The superior performance of the BERT model validates the hypothesis that transformer-based models, with their deeper understanding of context and language nuances, are well-suited for complex NLP tasks such as sarcasm detection. The high precision and recall indicate that BERT is effectively distinguishing between sarcastic and non-sarcastic classes, avoiding common pitfalls such as overfitting to specific linguistic cues that may not generalize well.

The lower performance of traditional models like Naive Bayes and Logistic Regression, despite their speed and simplicity, reinforces the need for more sophisticated approaches for tasks involving intricate language patterns.

The MLP's performance suggests that while deep learning models without transformer architecture can capture complexity to a degree, they may still fall short of the nuanced understanding that models like BERT provide.

In conclusion, the evaluation results underscore the importance of model selection in NLP and the potential of BERT for sarcasm detection. Future work could explore extending BERT's capabilities, perhaps by increasing the dataset size or incorporating additional context like user profiles or historical posting behavior to further improve performance. Additionally, cross-validation and hyperparameter tuning might refine the models' capabilities.

The report based on these findings would underscore the transformative impact of advanced models like BERT in the field of NLP and encourage the adoption of such models for tasks requiring a deep understanding of language.

References

[1] williamQyq, "GitHub - williamQyq/NNsonSarcasmDetection," *GitHub*. https://github.com/williamQyq/NNsonSarcasmDetection/tree/main

[2] "News headlines dataset for sarcasm detection," *Kaggle*, Jul. 03, 2019. https://www.kaggle.com/datasets/rmisra/news-headlines-dataset-for-sarcasm-detection

[3] Codertimo, "GitHub - codertimo/BERT-pytorch: Google AI 2018 BERT pytorch implementation," *GitHub*. https://github.com/codertimo/BERT-pytorch/tree/master

[4] R. Ali *et al.*, "Deep learning for sarcasm identification in news headlines," *Applied Sciences*, vol. 13, no. 9, p. 5586, Apr. 2023, doi: 10.3390/app13095586.

[5] C. Hashemi-Pour and B. Lutkevich, "BERT language model," *Enterprise AI*, Feb. 15, 2024. https://www.techtarget.com/searchenterpriseai/definition/BERT-language-model

[6] GfG, "Explanation of BERT Model NLP," *GeeksforGeeks*, Jan. 10, 2024. https://www.geeksforgeeks.org/explanation-of-bert-model-nlp/

[7] R. Misra and UC San Diego, "News headlines dataset for sarcasm detection." [Online]. Available: https://arxiv.org/pdf/2212.06035.pdf