Movie Reviews Sentiment Analysis System

An Automated ML End-to-End Pipeline

Hala Sedki
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
hala.sedki21d@eslsca.edu.eg

Yousef Hany
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
yousef.hany21d@eslsca.edu.eg

Laila Ayman
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
laila.abdelbaset21d@eslsca.edu.eg

Omar Hesham
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
omar.mohamed21d@eslsca.edu.eg

Hla El Sakka
Faculty of Computing & Digital Technologies
ESLSCA University
Giza, Egypt
hla.sakka21d@eslsca.edu.eg

Abstract — **Automated Machine Learning has become a popular method of improving the process of designing and implementing models in the context of machine learning. In this research paper, we describe the development of an efficient automated machine learning framework for sentiment analysis on IMDB dataset. There is nothing more important than knowing the polarity of textual data, and that is why the IMDB dataset is well suited for sentiment classification. Our pipeline includes careful data preparation, extensive model training and testing, careful assessment, and reliable deployment, and such advanced and proven methods.**

Keywords: AI, Speech, Emotions,

I. Introduction

Automated machine learning, known as AutoML, represents a real game-changer in how machine learning, from data preprocessing to model selection and hyperparameter tuning, is undertaken. With such an automatic approach, colossal interest has been given to make it easier to navigate through the bumpy roads of machine learning by dramatically decreasing the effort humans have to put in exploiting an algorithm. Even non-experts can then make use of them. This research paper will describe the construction of an end-to-end automated machine-learning pipeline for sentiment analysis on the IMDB dataset.

Sentiment analysis, at times, referred to as opinion mining, is a related natural language processing methodology for the determination and identification of the sentiment or subjectivity borne in a piece of text data. The relevance of this task is concerning product reviews, social media analysis, and customer feedback analysis. Essentially, with the correct form of sentiment analysis, you can verify and provide business minds with many public opinions to help drive business and organizational decision-making.

The IMDB dataset is a widely recognized benchmark dataset for sentiment analysis, consisting of a large collection of movie reviews labeled with positive or negative sentiments. The reviews themselves are packed with nuances of language, so sentiment analysis on this kind of dataset is always tricky and exciting. We move forward to build a robust, automated machine-learning pipeline to classify sentiments effectively, with a significant focus on the IMDB dataset.

Significant steps in our proposed pipeline are rigorous data pre-processing, which involves the transformation of raw text into some format amenable to a machine-learning model. Essential steps include tokenization, removal of stopping words, and vectorization using TF-IDF. Such pre-processing steps make sure that textual data has been converted into a numerical representation quickly for the induction of any machine-learning algorithms.

Then, we preprocess the data and train our models. Following logistic regression as a baseline, we further fine-tune the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model in the hope of further fine-tuning the model. Both models are then trained with the IMDB dataset and evaluated on the following standard metrics: accuracy, precision, recall, and F1 score. We can, therefore, deduce some of the strengths and weaknesses of such models by comparing how they perform in conducting such sentiment analysis tasks.

The MLOps practices we have implemented support considerable model saving, versioning, and deployments to ensure scalability and reproducibility. This will allow the integrating of the sentiment analysis model developed easily into practical applications and production environments.

The contributions of this research lie in the development of a comprehensive automated machine learning pipeline for sentiment analysis on the challenging IMDB dataset. Our pipeline leverages state-of-the-art techniques and practices to efficiently process the textual data, train models on the dataset, evaluate their performance, and deploy them for practical applications. The experimental results obtained through rigorous evaluation reveal that our proposed pipeline is efficient to employ, and the fine-tuned BERT model is significantly better than the logistic regression model. This paper presents our attempt to assist with the gap in AutoML that focuses on developing an efficient and autonomous model for sentiment analysis on the IMDB dataset, with the possibility of implementation in areas including, but not limited to, market research, customer feedback analysis, and social media analysis.

II. Related Works

Automated Machine Learning (AutoML) is one such concept that has gained much pride over the years in its rightful sense, and several studies have had their attention drawn to automating different stages of the machine learning pipeline. Hence, the section below provides a detailed summary of related work on AutoML, sentiment analysis, and the IMDB dataset.

Automated Machine Learning (AutoML):

Automated machine learning is geared toward reducing the much-needed, mainly manual, repetitive, and time-consuming processes of feature engineering, model selection, and hyperparameter tuning. Libraries and tools in this field, such as Auto-sklearn[1] H2O AutoML[2], and Google Cloud AutoML[3], are some of the most prominent behind making those efforts. These frameworks already provide pre-built solutions around topics that include automated model selection, optimization for hyperparameters, and ensembling techniques to ease the coaching process of coaching high-performing machine learning models. In principle, AutoML can also be applied in feature generation for any domain, including sentiment analysis. For example, in the work by Schafer and Chen [4], AutoML approaches to sentiment analysis using Twitter data were proposed. The pipeline is met by automated inferences of feature engineering techniques, like n-grams extraction, the use of sentiment lexicons, and automated model choice using a variety of classification algorithms.

Sentiment Analysis:

Sentiment Analysis, sometimes called opinion mining, is the classification of textual data into positive, negative, or neutral sentiment. There have been numerous ways to go about sentiment analysis over time, with these methods divided from classical machine learning to recent deep learning models. Among the popularly employed traditional machine learning methods in sentiment analysis are logistic regression, support vector

machines, and naive Bayes. A lot of data preprocessing is required by these methods toward data, feature selection, and model training for sentiment analysis [5].
Most of these algorithms are designed with handcrafted features and manual parameter tuning. Deep learning has been used more and more in the area of sentiment analysis, taking into consideration the capability of the recurrent neural networks (RNN) and convolutional neural networks (CNN) methods to model contextual information automatically [6] and learn features on their own. Recently, pre-trained contextualized word embeddings, such as BERT [7], show state-of-the-art results on most NLP benchmarks. Hailing from the Transformer architecture, BERT is a self-attention model that can relate words in a sentence to each other, therefore achieving state-of-the-art results on multiple benchmarks.

IMDB Dataset:

The IMDB dataset is one of the benchmark datasets for sentiment analysis. It includes movie reviews; for each review, there is a sentiment label from a set of two labels: either positive or negative. The dataset is balanced; hence, it makes it ideal for sentiment analysis per se, as it contains half the number of positive and negative reviews. The dataset is divided into 50,000 sentiments as positive and 50,000 sentiments as negative. The IMDB dataset has become one of the most established corpora in the development and benchmarking of models for sentiment analysis. The model-based recursive deep structure for sentiment analysis has also been tested and performs well in the IMDB dataset corpus. Besides, for some, more works have considered discussing multiple techniques, such as the ensemble methods [9], transfer learning [10], and domain adaptation [11], to enhance the sentiment polarity classification accuracy in the IMDB dataset. Another aspect seen is the proposal and testing of methods based on AutoML concerning the testing of the IMDB dataset toward the performance in the domain associated with sentiment analysis. Many researchers have compared the efficiency and effectiveness of several frameworks and algorithms of AutoML [4, 12].

Bert-Pretrained Model in Sentiment Analysis:

In contrast, such pre-trained models are pre-trained on massive textual data, capturing intricate semantic and contextual information to provide powerful representations for the downstream task.
There are reasonable instances of application of such models to sentiment analysis and similar projects. For example, the work demonstrates that, for pretraining on large corpses, sentiment classification accuracy is increased. Such embeddings capture broad semantic word relations that help the model understand the sentiment in the text much more effectively.
However, most importantly, the participation of transformer-based architectures, specifically BERT (Bidirectional Encoder Representations from Transformers), has proved to be one essential development for sentiment analysis with movie reviews. It is a pre-trained Google model on a huge range of different types of text data, like books and articles. It will be a step in the right direction if it has an attention mechanism, besides other special features like a bidirectional context model with rich semantic information.
Therefore, BERT has found application in sentiment-related tasks, such as movie review sentiment classification. Most literature fine-tunes BERT models with IMDB data and reports state-of-the-art benchmark performances.
For example, fine-tuning versions of BERT can set new state-of-the-art performance benchmarks on a range of NLP tasks, including sentiment analysis, well above previous methods.
Another necessary strength of BERT in sentiment analysis is its understanding of dependencies between words in a given context. The embedding in BERT is bidirectional, sufficiently capturing an all-in-context analysis of both preceding and succeeding words to make a clearer picture of the reflected sentiment in complex sentences. This contextual understanding infers more sentiment classification when the context of the sentence strongly drives the sentiment.
The BERT model can be further fine-tuned because of some particular sentiment analysis tasks in which

datasets are incredibly minute. That is, when a task-specific labeled, pre-trained BERT model like that of the IMDB dataset goes through its training, it is itself adapted to the sentiment analysis task. This further transfer learning category also possesses pretty cost-effective computation resources and training time while remaining pretty remarkable in performance.

Further research in this area has developed by working with and surpassing these studies using the pre-trained model, in most cases by BERT, for sentiment analysis in movie reviews. Alternatively, other approaches have either been based on handcrafted features or classical machine learning. Since BERT is contextual and can be fine-tuned to a specific task concerning sentiment analysis, it possibly turns out to be very useful, especially within movie reviews.

Conclusively, research in this sphere has not been left behind, given AutoML's important strides with the development of various models and tools in automating the multiple stages in the machine learning pipeline. There are a few significant tasks in natural language processing, and sentiment analysis is one of them. This is the only AutoML technique that is feasible to apply for automated feature engineering and model selection in natural language processing for sentiment analysis. Thus, much contribution has been derived from research based on the IMDB dataset for model assessment in sentiment analysis and, in this way, for comparison of the AutoML methods. The cocktail of AutoML, sentiment analysis, and datasets like IMDB is ripe for deep research and exploration.

III. Problem Statement

Classification of movie reviews for sentiment is a complex approach to the assessment of the linguistic moods prevailing in the public reaction to films and can be defined as the procedure of the automatic identification of the sentiment that is positive, negative or even neutral within the space of the text. Nevertheless, correctly identifying and categorizing sentiments in the movie reviews is challenging because of the use of such word forms, syntactic constructions, and personal opinions. Generally, traditional process that employs rule-based or handcrafted feature-based model can be unable to address the complex manner in which natural language and may have a tendency of missing the sentiments [5].

Another issue prevalent in sentiment analysis is the curation of large volumes of data which are annotated. When it comes to vast amount of data, manually labeling each review with the corresponding sentiment is a time consuming, costly, and impractical process hence restricting the scalability of the systems performing sentiment analysis [4].

In response to such challenges, the methods of pre-training the models and more so the transformer based models such as BERT have received considerable attention in the field. BERT has been emerged to have very high potential in terms of capturing contextual information and learn how to deal with the complexities of language due to the method of pretraining on a massive text of data [7].

However, more research is needed to enhance the efficient use of such models such as BERT for the sentiment analysis of movie reviews. Several questions may arise, such as the effect of further training the models with the movie review datasets, the different training techniques, and hyperparameters, and the comparison with other methods.

Due to this, the problem statement targets investigating the use of pretrained models to address the sentiment analysis of the given movie reviews, in this case, with technical bias. The first aim is to review the movie review datasets' applicability to fine-tuning BERT and assess its abilities in identifying sentiment correctly. This research shall entail the comparison between the performance of the BERT model when fine-tuned with the other approaches that include rule-based or handcrafted feature-based systems to determine the superiority of the fine-tuned BERT in sentiment nuance capture [12].

Moreover, the specific objectives of the work include the determination of proper training approaches and other

parameters to improve BERT fine-tuning for sentiment analysis. This investigation will entail comparing the effect of varying fine-tuning settings including learning rates, batch sizes and optimization algorithms to establish the optimal settings for the sentiment analysis tasks [10].

With regard to these technical issues, it aims to contribute to the development of the field of sentiment analysis and make a contribution to finding out whether the pretrained models are effective in sentiment classification or not. Hence, the findings of this research will be useful in shedding light on how to apply BERT in identifying the sentiment of movie reviews, and build more efficient and accurate sentiment analysis systems in the future. Such systems can successfully amalgamate and analyze the attitude of the audience towards the movies, which is helpful for film industries and production, marketing and recommendation systems.

IV. Methodology

This section introduces the method used as applied research. Then it previews sentiment analysis in movie reviews, first with traditional machine-learning techniques and ultimately with a proposed Language Model for pre-training BERT. This work describes the way data is collected, data pre-processing, exploratory data analysis, feature extraction works, model training along with evaluation, and steps followed in saving the fine-tuned BERT model for future use.

Data Collection and Preprocessing

Since it is a benchmark dataset, it is applied in most and even sentiment analysis[8]. The dataset includes movie comments with positive and negative sentiment expressed values, either having a positive or negative value. A custom-made function that load_IMDB_dataset is applied to load the dataset. This function reads the positive and negative sentiment labels [13] and extracts the corresponding text files to label the sentiment appropriately.

This is the preprocessed data for the sake of ease of the future analysis and modeling process. The function calls, and data preprocessing is done as follows: preprocess_data. The input review was tokenized using the BERT tokenizer. It normalizes the input sentence length by trimming the longer ones and padding the smaller ones with unique tokens. Attention masks are generated to indicate the presence of the tokens for the attention mechanisms [7].

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset and understand its characteristics. The distribution of sentiment labels was visualized using a count plot, providing an overview of the balance between positive and negative reviews. Additionally, the distribution of review lengths was examined through a histogram plot, enabling an understanding of the text length in the dataset. Word clouds were generated for positive and negative reviews to identify frequently occurring words [14].

Feature Extraction

Feature extraction was performed to capture relevant information from the cleaned reviews. Two types of features were extracted: text length and word count. Furthermore, the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique was employed using the TfidfVectorizer from scikit-learn [15]. This process converted the textual data into numerical representations and captured the importance of words within each review. Mutual information scores were calculated to select the top features based on their association with sentiment [16].

Model Training and Evaluation (Traditional Machine Learning)

The dataset was split into training and validation sets using the train_test_split method from scikit-learn. Logistic Regression, a widely used classifier, was trained on the extracted features and sentiment labels. The model was evaluated on the validation set using performance metrics such as accuracy, precision, recall, and F1 score. The logistic regression classifier was

chosen for its interpretability and the ability to handle high-dimensional feature spaces effectively [17].

BERT Fine-Tuning

We then just need to load the model and tokenizer, as we can work with pre-trained language models because of their powers. This is also done by the from_pretrained call from the transformers library. Next, we tokenized and encoded the dataset using the BERT tokenizer, and the tokenized dataset was then converted into PyTorch tensors and, finally, into a TensorDataset. The data loaders are created accordingly for the training and testing datasets.

The learning in this sense would be fine-tuning, that is, learning in some way. It inferred that it learned from the training data by applying a last layer of logistic regression on top of the BERT model. The loss function needs to be optimized by backpropagating, hence updating appropriate weights for fine-tuning. The test data is run on the final fine-tuned BERT model to get measures on different performance metrics like accuracy, precision, recall, and F1. By fine-tuning, it is possible to make the model capture fine-grained language representations and contextual information for precise sentiment analysis [7].

Model Saving

We save in the predefined path of where to save the fine-tuned BERT model using the predefined `save_pretrained` method of the tokenizer, hence similarly to below. This would enable saving the trained model and deploying the built automated end-to-end machine learning pipeline as a web interface, for instance, using tools like Flask and Docker.

Reproducibility and Research Transparency

An elaborate methodology ensures the reproducibility and transparency of research findings. The procedures for data collection, preprocessing, exploratory data analysis, feature extraction, and model training and evaluation could be detailed so that another researcher could attempt to replicate the findings, in turn building knowledge in the field and developing confidence in the researchers.

V. Experimental Results

The performance metrics of the model suggest it can correctly classify movie reviews with very high degrees of accuracy and effectiveness. The test accuracy given by the model came out to be about 0.8906, meaning that nearly 89% of the reviews were classified into positive or negative classes. This kind of accuracy makes a model strong and generalized very well on unseen data. The precision is 0.8837, meaning that for the number of optimistic predictions done, there is a very high ratio of truth. In other words, this is a high-precision model as it minimizes false positives while ensuring that most of the reviews predicted as positive were indeed positive. It is the recall score, in particular, hence the number of actual positives the model could determine correctly. The score should be 0.8995. It means that it is very good at revealing positive reviews, and the negative ones are very scantily named. The F1-score gives the harmonic mean of precision and recall to be 0.8915. This is quite a balanced score; that is, the model quite similarly describes both precision and recall, giving a satisfactory performance on all possible levels of classification.

This is effected by the learning model and is quite evident through the training process, as indicated by the loss values that progressively reduce with epochs. It starts from loss values of 0.3328 at the first epoch and trickles down to 0.0819 until the last one. This represents considerable improvements in performance with progressive stages of learning. In other words, the decrease of the graph with losses is almost linear, meaning the predictions become ever more right, passing through the training process. This is proved by the high evaluation metrics in solid measures. These results are evidential not only regarding the effectiveness of the approach taken but also regarding the power of the BERT model in carrying out this approach with preciseness in the capture and classification of sentiments in movie reviews.

VI. Conclusion

In conclusion, a benefit that is set up through sentiment-integrated analysis systems with the use of an automatic machine learning pipeline in the classification of textual data is very high test accuracy, which is at a value of 0.8906, excellent values for precision, recall, and F1, which fully validate the model developed to have passed through the actual intricacies of the sentiment analysis process. These metrics do not just illustrate the model's capability for correct predictions; they emphasize the robust reserve in minimizing mistakes while effective in balancing precision and recall.

Now that a pipeline of some sort is built up, advanced feature techniques help draw out the loading power of complex models like BERT. In this case, fine-tuning a pertained BERT model for detailed contextual information enabled sentiment classification to be carried out with reasonable accuracy. Research can be shown that an AMPL pipeline, under correct implementation, goes a long way to ease workflow processes and cut colossal traffic that the manual process was causing—lots of manual efforts in the process of training the models and tuning the hyperparameters. While an application and performance would be near to reality in this more significant scenario, this model has broad implications for large-scale market research, customer feedback analysis, and social media monitoring. In effect, automated sentiment analysis provides timely and accurate insights related to public opinion, fostering subsequent decision-making and strategy. Another significant upside is that since one of the properties of MLOps' practices is reproducibility and scalability in the pipeline, it gives more applicability to the model in real-world use. The results showed that fine-tuned pre-trained models outperform their traditional counterparts by providing superior performance for sentiment analysis tasks. It is well matched to tasks that hold many nuances in capturing the bidirectional context and dependencies between words, such as that presented here in the classification of movie reviews. Conclusion: This research contribution provides a proof of concept for an NLP automated machine-learning pipeline. Additionally, the performance metrics achieved by the fine-tuned BERT model still validate other evidence of its utility toward sentiment classification as a reliable and efficient tool for textual data analysis. In the future, there is still work to apply these optimization techniques to the pipeline in various domains to keep developing the frontier of automated sentiment analysis from a practical point of view.

VII. References

[1] Feurer, M., Klein, A., Eggensperger, K., Springenberg, J. T., Blum, M., & Hutter, F. (2015). Efficient and robust automated machine learning. In Advances in Neural Information Processing Systems (pp. 2962-2970).

[2] H2O.ai. (n.d.). H2O AutoML: Automated Machine Learning. Retrieved from https://www.h2o.ai/products/h2o-automl/

[3] Google Cloud AutoML. (n.d.). Retrieved from https://cloud.google.com/automl

[4] Schafer, L., & Chen, E. (2014). Automated sentiment analysis pipeline for Twitter data. In Proceedings of the 2014 International Conference on Data Science and Advanced Analytics (pp. 482-491). IEEE.

[5] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.

[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746-1751). Association for Computational Linguistics.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter

of the Association for Computational Linguistics (NAACL-HLT) (Vol. 1, pp. 4171-4186).
[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 Linguistics: Human Language Technologies (pp. 142-150). Association for Computational Linguistics.
[9] Wen, M., & Yang, D. (2018). Ensemble of deep learning architectures for sentiment classification. In Proceedings of the 2018 International Conference on Data Science and Advanced Analytics (pp. 619-628). IEEE.
[10] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL) (Vol. 1, pp. 328-339). Association for Computational Linguistics.
[11] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1253.
[12] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[13] IMDb Dataset. https://ai.stanford.edu/~amaas/data/sentiment/
[14] WordCloud. https://amueller.github.io/word_cloud/
[15] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.
[16] Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.