Redefining Words: The Simple Power of RCL in Text Classification

Eslam Ahmed Abdelrahman[*]

ABSTRACT

Abstract—In collaboration with Lumina AI, this paper highlights the effectiveness of the Lumina RCL model in text classification. We evaluate three diverse and challenging datasets, comparing Lumina RCL's performance against traditional machine learning and deep learning models. Our objective is to showcase Lumina RCL's strengths and potential in transforming text classification methodologies.

I. INTRODUCTION

The challenge with traditional machine learning is that it doesn't rise to complex tasks. The challenge with deep learning is that it requires extensive data, on one hand, and strong GPUs, on the other. In contrast to these, Random Contrast Learning (Lumina RCL™) represents a game-changing solution. It offers results like those of deep learning but no longer needs massive training data, and it has the cost-saving advantages of traditional machine learning. Its ability to achieve excellent performance with few resources and little data makes Lumina RCL a highly adaptable and flexible approach to text classification.

In 2022, Dr. Morten Middelfart, Sam Martin, and Ben Martin announced that they had created a new technique for machine learning called Random Contrast Learning (Lumina RCL). Lumina AI is the company behind this breakthrough. Established in 2015 in Tampa, Florida, Lumina AI is a specialized artificial intelligence company using its breakthroughs in AI and machine learning to develop CPU-based solutions for classification problems. Some theoretical comparisons will indicate the significance of their breakthrough.

In neural networks, any node in one layer is connected to several nodes in the next (fully connected layer). In contrast, Lumina RCL has a low number of connections; each node only has one parent in the next layer. Unlike the connectionist paradigm of neural networks, Lumina RCL is significant in terms of speed and is much more transparent. Traditional ML learns and trains like a human, and neural networks are designed to mimic the human brain in architecture. Lumina RCL was designed to imitate not the brain but the mind. The Lumina RCL model deviates significantly from traditional neural networks and from the transformer-structured models introduced in the famous paper "Attention is All You Need." While transformer models focus on the most important parts of a text, like how a spotlight works in image processing, the Lumina RCL model uses a different approach inspired by the philosophical tradition of Husserlian phenomenology. I will attempt to illustrate the difference.

Imagine you are looking at a painting. If we approach the experience from the perspective of brain science, we ask which neurons are firing and how we interpret the physical components of the environment before us. The approach of Husserlian phenomenology is not like this. Phenomenology abstracts from questions of natural science and modern epistemology and interrogates experience directly: What do I see, already interpreted, before me? What feelings already lie embedded within my perception? It attends to experience itself, not distracted by a mind-body problem. Neural networks process data with an architecture that imitates the brain. Lumina RCL imitates perception as it is

[*] Author Eslam Ahmed is a Machine Learning Engineer at Talents Arena, an AI and Machine Learning Instructor at more than one institution, e.g. Udacity, and an ML author at Educative. He holds two master's degrees in AI and ML from the University of Ottawa and IoT from Benha University. He is certified as a machine learning engineer and data scientist from Microsoft and IBM. Email: eslamahme@gmail.com

experienced, which is always already laden with stylized things, interpretations, and identities. It doesn't seek commonalities like text classification methods or NLP but instead proceeds immediately to find more interesting patterns in text, like finding hidden shapes in a painting.

In this research paper, we will present our RCL model testing and evaluation on three different domains of text classification data sets while comparing its results to machine learning models and neural networks.

II. DESCRIPTION OF DATASETS

The following datasets were used in this study and were obtained from public sources:

- Ecommerce Text Classification (50,425 samples, 4-class)
- Medical Text Dataset - Cancer Doc Classification (7,569 samples, 3-class)
- Customer Service chat Dataset (8,175 samples, 27-class)

2.1 E-commerce Text Classification Overview:

The E-commerce Text Dataset described four types of consumer goods - "Electronics", “Household”, “Books” and “Clothing & Accessories” in a dataset with one column for the class name and another for the data point of that class. With 50,425 texts included, the data point is the product plus description from all parts of an e-commerce website.

2.2 Medical Text Dataset -Cancer Doc Classification Overview:

The Medical Text Dataset Cancer Doc Classification focused on a long research paper to be classified into three classes: “Thyroid Cancer”, “Colon Cancer” and “Lung Cancer”. The Cancer Doc Classification Medical Text Dataset with 7,569 texts has two columns--one for class labels and one for the Research Paper Text of that class.

2.3 Customer Service Chat Dataset Overview:

The Customer Service Chat Dataset contains a total of 8,175 messages. The nature of inquiries for customer service chat messages is characterized into 27 categories. This involved two columns: "Message/Label" with "cancel order" or the like for what each heading describes, and "Chat Message" given as text.

III. METHODOLOGY

In this section, we explore the comparison and evaluation process of the Random Contrast Learning (Lumina RCL) model with traditional Machine Learning (ML) and neural networks. We provide a long and detailed exploration and discussion of Lumina RCL's unique features and operation techniques. We also describe how non-RCL experiments are carried out. Our purpose for this study is to highlight the unique features of Random Contrast Learning and its valuable and excellent text classification results.

3.1 Data Preprocessing and Preparation:

This section is about the key preprocessing and preparation steps for both Lumina RCL and non-RCL models, to guarantee that clean well-structured data are ready for use in analysis.

3.1.1.Data preprocessing:

For the non-RCL experiments:

- We used Label Encoder to transform text labels into numerical categories.
- We implement a custom function to clean text of HTML, URLs, accents, punctuation marks, non-alphabetic characters, excess whitespace, and stopwords.
- We dropped the duplicated records from the data to ensure that the dataset contains unique instances, preventing any potential biases or redundancies in the analysis
- Standard preparation includes segmenting the dataset into specified training and testing sets, with a testing ratio of 20%.
- For the non-RCL experiments across different models and datasets, text vectorization and word embedding were performed after dividing the data into training and testing sets. The details of these two operations will be discussed next in this context.

For the Lumina RCL experiments:

- We applied the same custom cleaning function and removed the duplicates as we did in the non-RCL experiments.

One of the crucial aspects of the Lumina RCL model is that it doesn't require extensive word embeddings or elaborate data preprocessing, as needed in the ML and DL experiments. Because of this feature, the Lumina RCL approach significantly reduces the amount of time in handling text data and in data preparation before training the model.

3.2 Data Preparation for RCL Experiments:

- **Text File Formatting:** Ensure each text sample is stored in a separate file.
- **Structure of Data Folder:** A primary folder contains many text files; these are then placed according to their categories into separate class-structured subfolders (such as **class1_folder, class2_folder**); be sure that all .txt files within these folders are unique.

3.3 Lumina RCL Model Experiment:

3.3.1 Overview of Random Contrast Learning (Lumina RCL) Model:

Lumina RCL attains higher performance and trains faster. This Lumina RCL model outperforms deep learning models in training speed, inference performance, and compact model size. The secret behind this is Lumina RCL’s use of a unique contrast technique that adapts the data in text classification by matching text samples with random text arrays, thereby filtering out non-unique patterns and illuminating patterns of greater specificity and interest.

3.3.2 Unique Methodology of Lumina RCL in Text Data Analysis:

The Lumina RCL model differentiates itself by comparing a text with a random text sample, rather than similar or dissimilar texts. It then adopts a model of filtering out coincidental patterns (which we do not really want) and focusing on both distinctive and interesting ones. With this approach, we grouped texts based on context-predicting word sequences and obtained an n-gram corpus that reflects the dataset's contextual flow. The differences between this method and regular deep learning models are obvious.

3.3.3 Experimental Setup and Technical Specifications:

In our experiment, we used the Lumina RCL Classifier (RCLC), via the PrismRCL application, on Windows machines powered by INTEL i7 processors with 128 GB RAM DDR4. The model's setup is designed to be simple, allowing optimization by way of a single command line. One of the most important features is the auto-optimized parameter within Lumina RCL which streamlines finding optimal parameters in training problems while working with the full CPU power in a multi-threaded application environment.

3.3.4 Repeated Experiments and Optimal Results Across Datasets:

The Lumina RCL and non-RCL experiments were applied to each dataset multiple times, but we will present only the best set of results achieved from each one.

3.4 Machine Learning Models Experiment:

This experiment was done on the same machine that we used for the Lumina RCL experiment: Intel i7 processors with 128 GB of RAM and DDR4, with CPU processing without a GPU.

TF-IDF vectorization:

We used the TF-IDF vectorization method to transform the text into a suitable format for these models.

Selected Machine Learning Models:

Over the same datasets, we used different machine learning models including Random Forest, SVM, Naïve Bayes, and Logistic Regression to compare the Lumina RCL model results with their results. This selection of models produced a wide range of results, showing the unique features and strengths of each model in text classification. We chose these models in a carefully systematic way covering a variety of technical methodologies, including tree-based and probabilistic approaches so that, we could do a comprehensive evaluation, contrasting Lumina RCL's performance with some of the most successful and high-performing models' architecture in text classification.

3.5 Deep Learning Models Experiment:

Neural network experiments were conducted on Google Colab, which is a cloud-based Python notebook environment with a lot of built-in support for machine learning. The Colab account used was configured with a T4 GPU, 15.4 GB of VRAM, and 12.7 GB of RAM.

Overview of Deep Learning Models Experiment:

In our deep learning model experiment with these three datasets, each dataset required a different architecture and schema depending on its specific characteristics and optimal performance needs.

To get the best fit for each dataset we carried out numerous tests before arriving at the best options. The architectures that we determined for each dataset are presented here.

- **E-commerce Text Classification and Medical Text -Cancer Doc datasets:**
  - Use **tokenization and padding** so that text sequences are converted into integer sequences of uniform length, with set maximum words and length parameters.
  - Build a sequential model with embedding, **convolution**, **max pooling**, **bidirectional LSTM**, **dropout**, and **dense** layers with different parameters in each data.
  - The model was compiled using **categorical cross-entropy** loss function and optimized with the **Adam** optimizer.
- **Customer Service Chat Dataset:**

  For the chat dataset, which is smaller than the others, we found that the GRU outperformed LSTM, and solving this set involved the following:
  - Use tokenization and padding so that text sequences are converted into integer sequences of uniform length, using the **FastText** model for vectorization with a designated vector size.
  - Built a sequential model with embedding, **GRU**, **dropout**, and **dense** layers but different parameters for each dataset.
  - The model was compiled using **sparse categorical cross-entropy** loss function and optimized with the **Adam** optimizer.

IV. RESULTS

The detailed results presented in this section show the performance of Lumina RCL in comparison with other machine learning and neural network models. They also exhibit Lumina RCL's unique advantages in accuracy and in its defense against data leakage when data contains duplicates or is too small to be trained. Lumina RCL provides metrics for recall, fl-score, accuracy, precision for each class, an overall accuracy for all classes. Consequently, accuracy has been chosen as the primary metric in our comparison of model performance. It will demonstrate Lumina RCL’s unique capabilities and features.

4.1 Performance Metrics of Lumina RCL vs. Non-RCL Models:

4.1.1 Non-RCL Results:

According to non-RCL experiments results, the traditional machine learning models, and neural networks both have similar levels of accuracy, as shown in Table 4.1. The main difference between them lies in training time duration. For example, models like the Random Forest and SVM require significant training times. Experiments with the E-commerce dataset show a trade-off between accuracy and time efficiency.

Data	Samples	No. of classes	Value	Random forest	Naïve Bayes	Logistic regression	SVM	Neural network
E-commerce	50,425	4	Accuracy	93.4%	93.3%	94.7%	95.3%	95.4%
			Time (sec)	199.679	1.186	6.648	283.407	573
Cancer-Doc	7,569	3	Accuracy	66.6%	43.2%	72.7%	70.7%	73%
			Time (sec)	8.607	1.652	4.15	8.104	71
Customer Chat	8,175	27	Accuracy	98.3%	98.7%	98.8%	99%	97.9%
			Time (sec)	2.298	0.049	0.342	0.905	19.8

Table 4.1. Accuracy values and training times for the different methods of machine learning and deep learning employed without duplicates

The Neural Network model always takes more than double the training time across datasets with the same level of accuracy, as is clear in the E-commerce and Customer Chat datasets. A considerable amount of time is needed for training these models. Note that the Naïve Bayes model has both the shortest training time among all models and the worst accuracy. These experiments demonstrate the trade-off between training time and accuracy: While Neural Networks can achieve high accuracy, they incur significant costs in increasing training time.

4.1.2 Lumina RCL Results:

In our Lumina RCL model results, Lumina RCL exhibits a particularly high level of performance across datasets with varied complexity. The Lumina RCL model results for text classification tasks show a remarkable balance, with similar levels of accuracy and training speed. See the comparison with other models in Table 4.2. Experiments with the E-commerce dataset showed its filing capabilities for 93.8% accuracy in just over 2 seconds, as well as its excellent performance with 27 classes in Customer Chat; it achieved 99.1% accuracy in only 0.156 seconds. Experiments with the Cancer-Doc dataset also demonstrated how quick Lumina RCL is at handling data: it achieved 78% accuracy in about 4 seconds. The data shows Lumina RCL's expertise in dealing with different class complexities quickly, while maintaining accuracy similar to its most quickly trained competitors.

Data	Samples	No. of classes	Accuracy	Time
E-commerce	50,425	4	93.8%	2.094
Cancer-Doc	7,569	3	78.0%	4.125
Customer Chat	8,175	27	99.1%	0.156

Table 4.2. RCL's accuracy values and training times for the given numbers of classes and samples without duplicates

4.2 Data Duplication Experiment:

In this part, we will introduce some experiments that we ran to measure the quality of different text classification models under data duplication, which is a common type of data leakage.

Data	RCL	Random forest	Naïve Bayes	Logistic regression	SVM	Neural network
E-commerce	95.2%	97.7%	94.8%	97%	98%	98.04%
Cancer-Doc	88.4%	100%	92.4%	94.9%	92.6%	96%
Customer Chat	99.6%	98.8%	99.6%	99.4%	99.2%	98.5%

Table 4.3. Accuracy values for RCL versus Non-RCL models with duplicated data

Data leakage can significantly impact a model's performance metrics; for example, duplicates in the data may cause bias or overfitting effects.

Data	RCL	Random forest	Naïve Bayes	Logistic regression	SVM	Neural network
E-commerce	0.014	0.043	0.015	0.023	0.027	0.0264
Cancer-Doc	0.104	0.334	0.492	0.222	0.219	0.23
Customer Chat	0.005	0.005	0.009	0.006	0.002	0.006

Table 4.4 Differences between accuracy values with and without duplicates in the data

Our objective was to know how each model's performance metrics would change after removing duplicates from the data. Table 4.4 illustrates how data deduplication affects model accuracy: We compare the performance with duplicated data, as shown in Table 4.3, with the performance after removing duplicates in Table 4.1 and Table 4.2. Table 4.4 reports the difference between them (delta value = accuracy with duplicates – accuracy after removing duplicates). The robust Lumina RCL model has only 0.014 of variance in accuracy before and after duplicates for E-commerce data, 0.104 for Cancer-Doc and 0.005 Customer Chat data. This shows that there is little impact from data duplicates, which stands in great contrast to other models like Random Forest and Naïve Bayes, whose data shifts are 0.043 and 0.492 in E-commerce. The differences highlight the data leakage that affects traditional models and how much more reliable the Lumina RCL model is in contrast. The following delta charts for each dataset better illustrate the point: Lumina RCL outperforms in producing dependable measure and in adapting to data leakage.

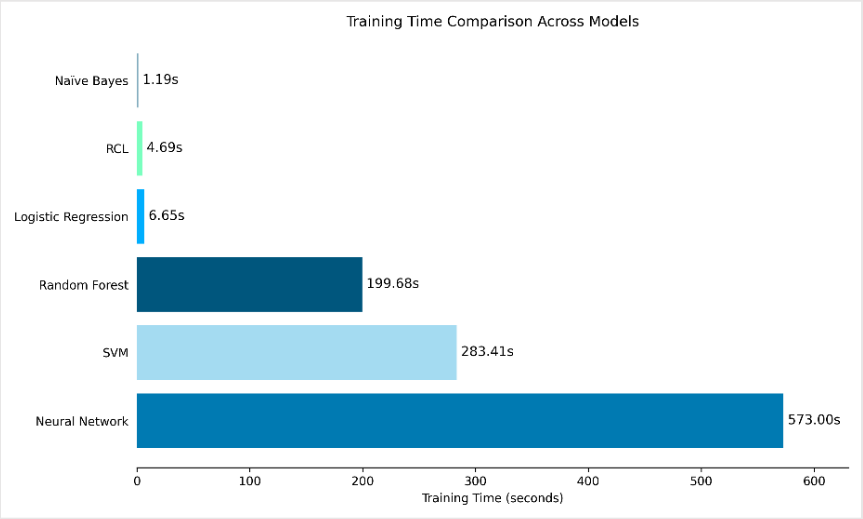

Figure 4.1. Training Time for E-commerce data set to compare between Lumina RCL and non-RCL models

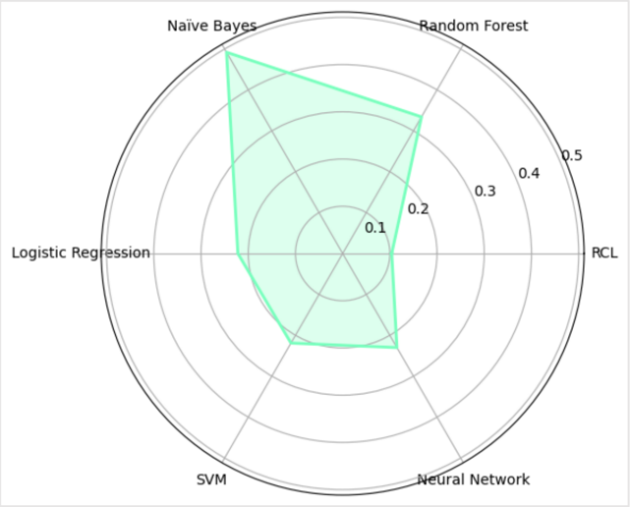

Figure 4.2. Cancer documentation delta values (the difference between accuracy after and before removing duplicates)

4. Overall Comparative Analysis:

In the comparative analysis of various datasets, Random Contrast Learning (Lumina RCL) demonstrated remarkable efficiency. Although it had an accuracy equivalent to other advanced models, it took a much shorter time to train than others. Figure 4.1 shows how much more quickly Lumina RCL trains in comparison to larger models like Neural Networks. Despite their high accuracy, the latter demand longer training periods. Note that while Naive Bayes is close to Lumina RCL in training time, it didn’t reach Lumina RCL's accuracy and was the most affected by data leakages and overfitting, as shown in Figure 4.2. This shows that Lumina RCL has a very structured performance, i.e. within the same level of accuracy it is capable of better performance, regardless of how much complexity results from data leaks and overfitting. This demonstrates its practical feasibility for real scenarios that require both accuracy and speed.

In the challenging case of the Medical Text Dataset - Cancer Doc Classification, deduplication reduced the dataset from 7,569 records to just 996 records, and model performance stability was challenged, as shown in Figure 4.2 and Figure 4.3 below. The Lumina RCL model provided better accuracy with minimal impact from data duplication. Lumina RCL exhibited extremely robust performance on the Cancer Doc dataset despite massive data reduction. This is enough to confirm its worth for small datasets, in addition to its avoidance of overfitting or underfitting as affect other models. This consistent performance entails that the Lumina RCL model is better designed for learning from datasets that suffer from potential leakage or have limited training data.

Data	Value	Random forest	Naïve Bayes	Logistic regression	SVM	Neural network	RCL
E-commerce	Accuracy	93.4%	93.3%	94.7%	95.3%	95.4%	93.8%
	Time (sec)	199.679	1.186	6.648	283.407	573	2.094
Cancer-Doc	Accuracy	66.6%	43.2%	72.7%	70.7%	73%	78.0%
	Time (sec)	8.607	1.652	4.15	8.104	71	4.125
Customer Chat	Accuracy	98.3%	98.7%	98.8%	99%	97.9%	99.1%
	Time (sec)	2.298	0.049	0.342	0.905	19.8	0.156

Table 4.5. Accuracy and Time values for Lumina RCL versus Non-RCL models after removing duplicates

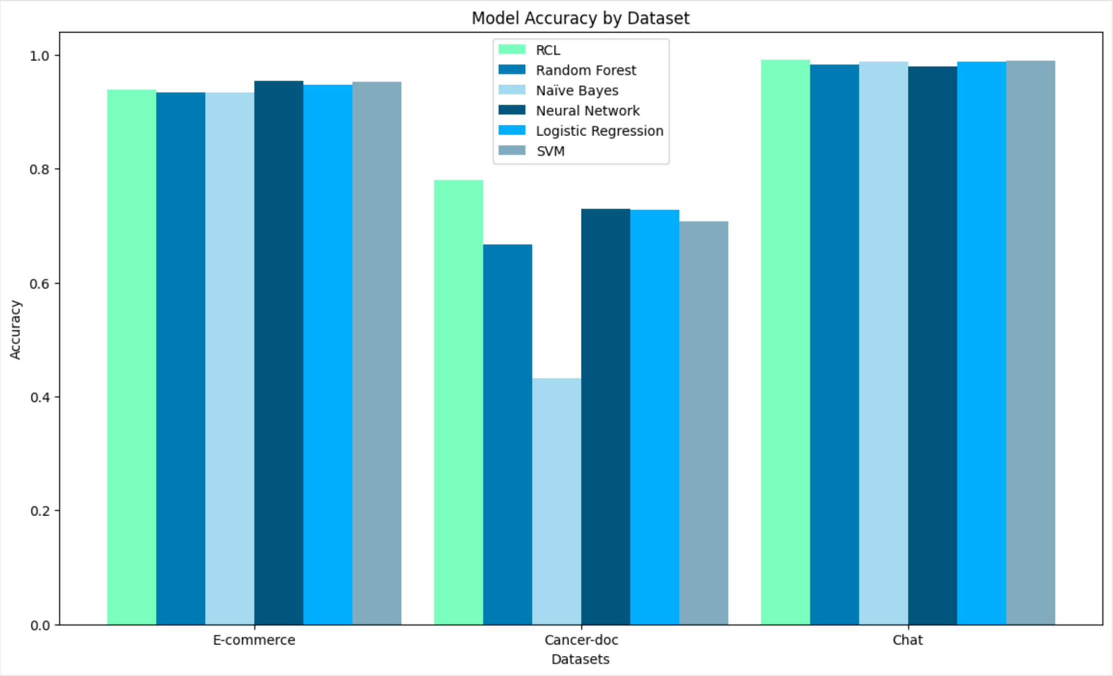

Figure 4.3. Accuracy comparison between Lumina RCL and non-RCL models for the three datasets

V. CONCLUSION

This research paper evaluates how well the Lumina RCL model performs against traditional machine learning and deep learning in text classification. In our research, we experimented with three independent datasets, each presenting a different classification challenge. The Lumina RCL model, a newly emerging player on the scene of machine learning, was distinguished by a curious mixture of speed in training and high accuracy. It stood out in contrast with traditional methods which tend towards one or the other of these advantages. The Lumina RCL model's efficiency did not lead to lower accuracy; instead, it consistently produced results comparable to, if not superior to, those of other models. We focused on data leakage and overfitting effects, especially due to duplicate data, and how this impact model performance. The Lumina RCL model withstood this challenge. Although results across models were similar on most data, Lumina RCL stood out as preferable in cases of limited training data. In conclusion, the Lumina RCL model proves a highly effective tool for text classification. Lumina RCL impressively combines the power of deep learning with the practicality of ML approaches.