Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 6, No. 2, September 2023, pp. 131 – 143
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Hybrid Machine Learning Techniques for Comparative Opinion Mining

Bernard Ondara[1], Stephen Waithaka[2], John Kandiri[3], Lawrence Muchemi[4]
[1,2,3,4]Department of Computing and Information Technology, Kenyatta University, Kenya
[3]Department of Computing and Informatics, University of Nairobi, Kenya
Email: [1]ondara.bernard@ku.ac.ke, [2]waithaka.stephen@ku.ac.ke,
[3]kandiri.John@ku.ac.ke, [4]lmuchemi@uonbi.ac.ke

Article history:
Received May 10th, 2023
Revised Jun 26th, 2023
Accepted Aug 6th, 2023

Keyword:
Classification
Data Sharing
Electronic Health Record
Radial Basis Function
Support Vector Machine

ABSTRACT

Comparative opinion mining has lately gained traction among individuals and businesses due to its growing range of applications in brand reputation monitoring and consumer decision making among others. Past research in sub-field of opinion mining have mostly explored single-entity opinion mining models and the mining of comparative sentences suing single classifiers. Most of these studies relied on a limited number of comparative opinion labels and datasets while applying the techniques in limited domains. Consequently, the reported performances of the techniques might not be optimal in some cases like working with big data. In this study, however, we developed four hybrid machine learning techniques, with which we performed multi-class based comparative opinion mining using three datasets from different domains. From our results, the best-performing hybrid machine learning technique for comparative opinion mining using a multi-layer perceptron as the base estimator was the Multilayer Perceptron + Random Forest (MLP + RF). This technique had an average accuracy of 93.0% and an F1-score of 93.0%. These results show that our hybrid machine learning techniques could reliably be used for comparative opinion mining to support business needs like brand reputation monitoring.

Corresponding Author:
Bernard Ondara,
Department of Computing and Information Technology,
Kenyatta University,
P.O. Box 43844-00100 Nairobi, Kenya.
Email: ondara.bernard@ku.ac.ke

DOI: http://dx.doi.org/10.24014/ijaidm.v6i2.22644

1. INTRODUCTION

Huge volumes of user-generated content are created daily on popular social media platforms such as Twitter and Facebook [1]. The application of opinion mining on such platforms is often aimed at analyzing user sentiments toward specific entities. Comparative opinion mining is a distinct type of opinion mining where the focus is on extracting and classifying opinions from user-generated content where comparisons between different entities are involved. Examples of these entities include brands, products, services, politicians, and governments [2]. To date, comparative opinion mining tasks have mainly been performed on English and Chinese datasets [3].

A useful source of comparative opinion mining data is product reviews [4], which forms the basis of many prospective customers choosing their preferred entities (brands, products, et cetera) [5]. In this case, comparative reviews must be distinguished from regular reviews. In regular reviews, one is interested in the opinions targeted at a specific entity regardless of any comparisons made between mentioned entities. Such opinions are also known as direct opinions [2]. Conversely, comparative opinion mining is premised on extracting sentiments from opinions that target multiple comparable entities. Because online reviews are often written in an informal language, performing comparative opinion mining using such reviews is not a

trivial task [6]. Recently, one of the richest environments for consumers to exchange opinions about products is product discussion forums. According to [7], [8], and [9], from these forums, users can choose their preferred products or brands, hence making their decision making easier.

Comparative opinion mining is helpful to brands in as far as trying to gauge the sentiments that their customers have towards their specific business solutions (products and/or services). An example of a comparative statement or sentence is: "Nokia phones are more durable than Samsung phones". This sentence compares Nokia phones with Samsung phones based on the durability aspect of the two brands' smartphones. In this example, the user (opinion holder) prefers Nokia phones (positive sentiment) yet Samsung would benefit from knowing that their customer has an issue with the durability of their phones, as depicted in the negative sentiment. This may explain why some customers go for Nokia phones. Comparative opinion mining tools are essential for potential customers who need information about competitor products before buying a product/service from a specific brand [6].

Early research in comparative opinion mining using the machine learning approach predominantly relied on limited datasets and a small set of different machine learning classification algorithms [10]. Therefore, it is paramount to attempt the application of hybrid machine learning techniques on bigger datasets from popular data sources to determine and consequently recommend the most optimal hybrid machine learning technique for performing comparative opinion mining. The development of hybrid machine learning techniques for comparative opinion mining in this study was motivated by research findings showing that hybrid techniques have improved performance, enhance model generalization and robustness [11], have better transfer learning and domain adaptation [12], and better address data sparsity [13] considering the relatively small sizes of the datasets used, and the possibility of handling additional sentiment aspects. Our baseline studies included those by [14] that applied single-entity classifiers on a limited dataset resulting in poor performance and [15] that developed a hybrid machine-learning technique by fusing an Artificial Neural Network (ANN) with a Decision Tree and applied it to energy consumption prediction.

In this study, we included a deep learning technique in our hybrid machine learning models. According to [16] and [17], deep learning techniques perform better than traditional machine learning techniques on large datasets like those from social media platforms. Research shows that the size of the dataset, number of features, feature-extraction techniques, and application domains affect machine learning algorithm performance and the quality of results [16]. To overcome most of these challenges, our proposed hybrid techniques perform comparative opinion mining on relatively bigger datasets, considering that comparative opinions exist in about 10% of user generated content [2]. Our experiments involved a deep learning technique as a base estimator and a traditional machine learning technique as the final estimator. The use of transformers has been tried out in the recent past, demonstrating their ability to extract textual dependencies, handle many languages, support transfer learning, and make it possible to perform tend-to-end learning. However, their use is faced with limited model interpretability and high computational requirements [18], and reliance on huge labelled datasets [19]. Obtaining huge labelled datasets to support transfer learning, for instance, requires a lot of time and is costly. For these reasons, this study could not use transformers. The Count Vectorizer feature engineering technique was used to help contextualize the words being analyzed [20]. This is important in extracting the opinions towards each entity in the user-generated content [16]. The additional word contextualization was addressed by using higher-order n-grams, in this case trigrams [21].

Research Questions

RQ1: How would hybrid machine learning techniques for comparative opinion mining be developed?
RQ2: How would hybrid machine learning techniques for comparative opinion mining be evaluated?
RQ3: What is the performance of the hybrid machine learning techniques in comparative opinion mining?
RQ4: Which one is the most efficient hybrid machine learning technique for comparative opinion mining?
RQ5: Would hybrid machine learning techniques be reliable for comparative opinion mining?

Research Contributions

To carry out comparative opinion mining using hybrid machine learning techniques for applications such as brand reputation monitoring using online reviews, this research presents multiple significant contributions:

1. The development of four hybrid machine learning techniques for comparative opinion mining.
2. Applying the developed hybrid machine learning techniques offers a basis for opinion mining researchers to design much more complex hybrid machine learning techniques for comparative opinion mining.

3. The performance of the hybrid machine learning techniques in this study will help future researchers in selecting optimal machine learning techniques that could be hybridized for comparative opinion mining.
4. Our findings show that hybrid machine learning techniques can be reliably used in comparative opinion mining.

Research by [22] used two supervised algorithms and three greedy algorithms and obtained satisfactory results [22] through optimization. However, their results were based on three different datasets. The kind or nature of the dataset affects the performance of an algorithm. This is a limitation of this study because varying datasets yields varying results. Other studies on comparative opinion mining have been done on movie reviews [23] where Synthetic Word, Linear Support Vector Machine, and Naive Bayes techniques were applied using the standard IMDB dataset. In this study, the best-performing algorithm was the Linear SVM and the system they proposed could work with many datasets from various domains. [24] proposed a system for comparative opinion mining applied to Twitter reviews while [25] used movie reviews from Twitter data showing that SVM had the best results.

A study was done by [2] in which they attempted SVM and a clustering technique with a manually generated dataset from Amazon for product reviews. In their findings, SVM produced 79.8% accuracy. Another study by [26] also used n-grams to test different approaches to performing opinion mining on Amazon Product Reviews. However, this study is limited by the exploitation of a small set of machine learning classifiers. A different study on hybrid machine learning algorithms proposed by [2] merged SVM with clustering. The drawback of this study is the application of the model to one domain. Better results could be obtained by experimenting with many domains. A study by [16] proposed a system that uses comparative as well as superlative types of sentences using datasets from Amazon, Howard forums, and CNET. From these studies, a common challenge observed is that of obtaining results using generalized comparative opinion mining systems that do not depend on training datasets.

While deep learning techniques have been used successfully in direct-opinion mining such as in the studies by [27], [17], and [28] there is scanty research on the development of hybrid machine learning classifiers that include a deep learning technique in the architecture. According to [16], traditional machine learning approaches have been used in subjectivity classification [29], and opinion detection. This research shows, however, that deep learning approaches are not common in subjectivity classification. Moreover, [16] found that hybrid approaches were not common in subjectivity classification as well. These are areas of research that need attention. While using deep learning techniques in opinion mining where datasets are big would offer performance gains, the performance must be evaluated across multiple domains. This, according to [16] would avoid generalizing the performance of a technique. Rather, averaging the performance across multiple datasets gives a better indication of the performance of a technique. This is because a technique may perform well in one domain but poorly in a different domain.

Table 1: Research Works Related to This Study

Study	Aim	ML Techniques	Datasets	Performance	Limitations / Our Recommendations
Varathan et al. [2]	Comparative Opinion Mining	SVM Clustering	Compiled from Amazon	79.8% for SVM	Only one domain of product reviews was used.
Khan et al. [9]	Comparative Opinion Mining	NB	YouTube User Reviews (iPhone and Android)	33%	Poor performance. Multiple classifiers are needed for better performance.
Bhavitha et al. [18]	Sentiment Analysis of Comparative Reviews	SVM	Tech Product Reviews; Movie Reviews	85%	No Automatic Analysis. Only SVM was experimented with.
Tkachenko & Lauw [30]	Entity Comparisons	SVM	Amazon Reviews	Better performance than past baselines	Could be improved by using multiple datasets and trying different classifiers
Younis et al. [10]	Comparative Opinion Mining of Online Reviews	NB,LR, SVM, KNN, DT, RF, GB	Online Reviews from Kaggle.com	RF accuracy of 95% and F1-score of 95%.	Experimenting with Hybrid Techniques involving deep learning to improve prediction on large datasets is necessary.

The above works show significant progress in comparative opinion mining. However, there are some gaps that call for further research. For instance, [25] used one dataset, [2] used one domain, [18] used one machine learning technique, while the technique by [9] had below average performance. Furthermore, [13] and [23] observes that there is minimal progress in developing hybrid techniques for subjectivity

classification. Likewise, deep learning techniques have seldom been used in developing hybrid architectures [23]. To overcome these challenges, our study developed hybrid machine learning techniques containing both a deep learning and a traditional machine learning technique. The hybrid techniques were tested across multiple dataset and domains. Our results reveal improved performance over single classifiers. This is partly due to hybrid techniques addressing data sparsity in a better way [12]. The hybrid techniques also offer improved extraction of additional sentiment aspects.

2. RESEARCH METHOD

The hybrid machine learning techniques for comparative opinion mining proposed in this work were developed through the following stages: (1) Collection of Annotated Data, (2) Pre-processing of data, (3) Development of Hybrid Machine Learning Classifiers, (4) Application of the hybrid machine learning classifiers on online comparative reviews, (5) Performance evaluation of the various hybrid machine learning techniques, and (6) determination and recommendation of the most optimal technique for application in comparative opinion mining. In this method, we combined two independent techniques: a deep learning technique as the base estimator and a traditional machine learning technique as the final estimator. This was done to leverage the power of each technique while minimizing the limitations of each technique that forms part of the hybrid technique thereby improving the overall performance on large and imbalanced datasets [31].

2.1. Dataset Collection & Annotation

The datasets used in this study consisted of data containing comparative reviews. We collected three datasets, which belong to three application domains: (1) technology brands represented by the Microsoft vs Google dataset, (2) social media brands represented by Facebook vs Twitter, and (3) smartphone operating systems brands represented by iOS vs Android. A more detailed description of these three datasets is shown in Table 2 below. The sentences were pre-annotated according to their sentiments: positive, neutral, and negative. These classes are further divided into nine classes: pos_neg, pos_pos, pos_neu, neg_pos, neg_neg, neg_neu, neu_neu, neu_neg, and neu_pos. The datasets were downloaded from https://www.kaggle.com/umairyounis/comparative-reviews-datasets. The datasets obtained were split into two: (1) 80% of the data as training data, and (2) 20% of the data as testing data.

Table 2: Details of Datasets Used

Dataset	Reviews	pos_pos	pos_neg	pos_neu	neg_neg	neg_pos	neg_neu	neu_neu	neu_pos	neu_neg
Microsoft vs Google	3011	360	1268	396	62	380	46	321	148	30
Facebook vs Twitter	3000	440	1208	447	54	307	59	310	143	32
Pearl Continental vs Marriott	1012	276	138	46	92	138	46	138	92	46

2.2. Training Data

Three human annotators took part in manually verifying the assigned sentiment polarity classes to the training dataset. To come up with the final polarity, the most common polarities assigned to each review by the three annotators were applied. This concept is equivalent to majority votes in an election. The agreement percentage among the annotators was 82.7%, which is satisfactory. The Kappa (K) score was 0.81, which shows there was strong agreement among the human annotators. Table 3 below presents a sample of the training dataset. The percentage agreement level was obtained by dividing 5,808 (the number of times the annotators assigned the same class to a review) by 7,023 (the total number of reviews that were annotated).

2.3. Testing Data

A testing set is used to evaluate the model by serving as a benchmark. The test data is applied after the model has been fully trained on the training data and this happens at the classifier testing stage. Through model validation using the test data, one can determine if the model is working or not [22], [32]. A random split method with an input of the 80%: 20% ratio was used to split the dataset into two categories: training dataset and testing dataset. This method is better than other methods and therefore yields more accurately partitioned datasets [5]. The datasets were stored in CSV files for purposes of experimentation. The obtained datasets were cleansed and processed before they were fed into the various hybrid machine learning classifiers [24].

2.4. Data Preprocessing

Before preprocessing the data, the datasets were cleaned in the following ways for consistency.

1. Any whitespaces in the column names were removed to ensure uniformity and for easy replication in later steps.
2. White spaces in the label names were also removed since the analysis would treat two labels with the same name but with extra whitespace as different entities. Besides the removal of unimportant special characters, and parts of speech (POS) tagging, the following tasks were performed.
   a. **Tokenization -** This is the process through which the words in the collected data are broken down into small chunks of text. To achieve this, we used NLTK Tokenizer in Python**.**
   b. **Stop Words Elimination -** In user-generated content, not all words carry opinions. Such words are called stop words. To achieve this, we used a Python script with a set of a pre-defined list of stop words, which include but are not limited to "a", "an", "the", and "is".

2.5. Applying Hybrid Machine Learning Techniques

In this study, we experimented with four hybrid machine learning techniques for performing comparative opinion mining. These techniques included MLP_DT, MLP_RF, MLP_SGD, and MLP_SVM. The idea behind these hybrid techniques was to use a deep learning technique as the base estimator and a traditional machine learning classifier as the top-level (final) estimator. During this stage, we used the partitioned datasets, which consist of training sets and test sets. The reviews are labeled according to their classes: positive, negative, and neutral for each brand entity mentioned in the data. Figure 1 below shows a generic illustration of a supervised machine-learning technique.

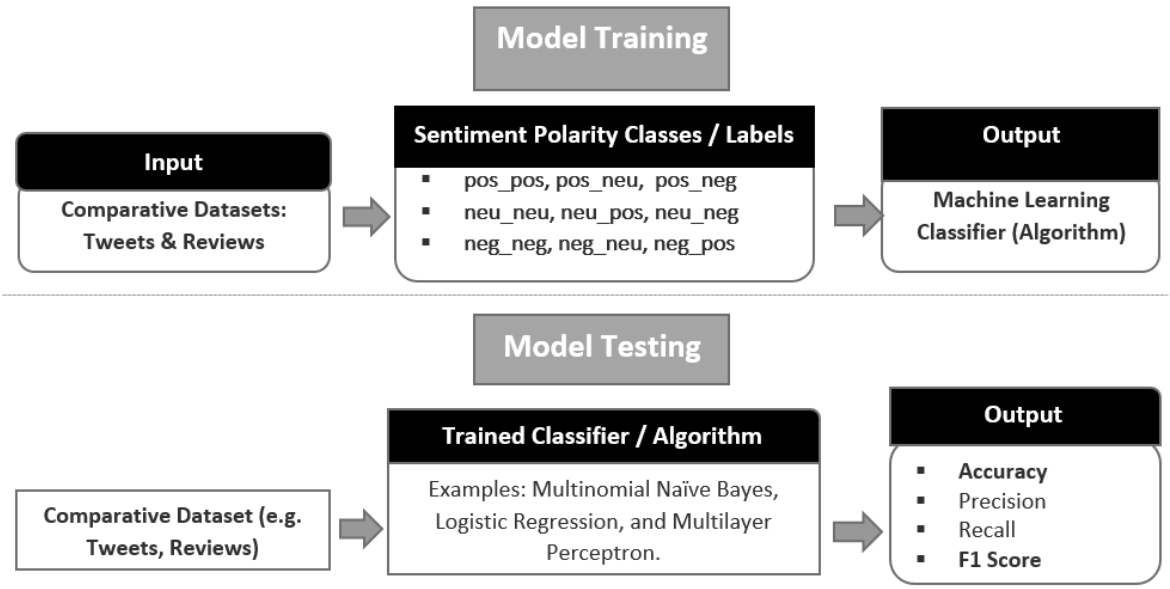

Figure 1. Machine Learning-based Process for Comparative Opinion Mining.

2.5.1. Feature Engineering

To implement the hybrid machine learning techniques in our study, the following feature selection steps were adopted: (1) Feature Vector, and (2) Term Frequency (TF).

1. Feature Vector - converts a review into a specific matrix of token counts [33]
2. Term Frequency (TF) - counts the frequency of a term in the provided reviews [33].

We used the Bag of Words (BOW) method with n-grams (the counts of words) [16] with a range of 3 (i.e. trigrams) to produce higher classification accuracy with the context of terms (words) in mind [16]. A study by [34] found that hybrid methods achieved the best CPU performance and accuracy levels even on feature selection. In the subsequent sub-sections, we summarize the machine learning techniques and a popular deep learning technique we used in developing our hybrid machine learning techniques for experimentation.

2.5.2. Multi-Layer Perceptron (MLP)

A Multilayer Perceptron (MLP) algorithm is used to supplement the feedforward neural network. One of the use cases of MLP is classification and prediction. In this model, the output layer, which is responsible for classification receives input from the input layers. Due to the feed-forward nature of the MLP network, data flows forward from the input layers to the output layer. However, the backpropagation learning algorithm is used to train the neurons making up the MLP architecture. Each node in the MLP architecture makes use of a sigmoid activation function shown below to accept an input of real values and transform such values into a number ranging from 0 to 1. Research by [35] demonstrates that MLP can achieve high accuracies in classification.

$$\alpha(x) = 1/(1 + exp(-x)) \tag{1}$$

2.5.3. Decision Tree (DT)

This is a supervised machine learning technique that is often useful in solving text classification problems. Feature values are sorted before they are used to classify instances in the text. Each node in the decision tree is constituted by a node while feature values are represented as branches of the tree. By sorting the values of features, the classification is accomplished starting from the root node of the decision tree. This technique implements a divide-and-conquer technique to construct the decision tree[24]. To represent a decision tree using a mathematical formula or expression, assume ***a*** represents the root node while ***b*** represents a subset, and ***x*** represents the leaves of the tree, then, the following mathematical formula is constituted.

$$(a, b) = (x_1, x_2, x_{3,} x_{4,} x_5 \ldots x_{k,\, b}) \tag{2}$$

2.5.4. Random Forest (RF)

The Random Forest algorithm is considered an easy-to-apply machine learning classifier that involves tuning hyper-parameters. It is also reputed as the most flexible machine learning classifier in addition to most frequently producing efficient results. Constructed from several decision trees, the RF algorithm is an ensemble algorithm of many decision trees. If the RF has more decision trees, it will have improved results because of enhanced generalizations [14]. The RF technique could be represented using the following mathematical notation:

$$y = \frac{1}{z} \sum_{z=1}^{x} yz(x') \tag{3}$$

where ***1*** denotes the total of samples, ***Z*** represents the number of training instances from ***x, y*** while ***2*** denotes training a classification tree represented by ***yz***.

2.5.5. Stochastic Gradient Descent

This popular machine learning technique is iterative and often used for purposes of making an objective function optimal though appropriate smoothness properties such as sub-differentiable and differentiable properties. It is the foundation of neural networks. The logic behind this algorithm is that it begins from some random point in a function and moves down its slow step by step till the lowest point on that function**.** Given that SGD picks at random a data point from a data set during every iteration, this technique achieves enormous computation reduction [36]. Thus, SGD is efficient in text classification. Still, the performance degrades with increasing dataset sizes.

2.5.6. Support Vector Machine (SVM)

This is a supervised machine learning technique that handles text classification tasks by sorting the given data into different classes based on the discovery of a hyperplane (line) for splitting the dataset into many classes [37]. In our case, the documents or corpus containing the data to be classified contains reviews. This technique can be mathematically represented as follows:

$$D = \{(x1, y1), (x2, y2), \ldots\ldots (xy, yn)\} \tag{4}$$

where ***D*** denotes the dataset and ***x*** and ***y*** are variables showing the relationship between them for class elements.

2.6. The Proposed Hybrid Machine Learning Technique for Comparative Opinion Mining of Online Comparative Opinion Reviews

The supervised machine learning technique applied to comparative opinion mining in our study work by accepting as their inputs, reviews, then applying the various data pre-processing steps outlined earlier before finally classifying the data into various polarity-based classes. The classes used in our experiment were positive, negative, and neutral.

The dataset was split in the ratio of 80% to 20%, being the training set and testing set respectively. At the training stage, the machine learning classifier is fed with the polarity label and the review itself. After model training, the test dataset is used to evaluate the efficiency of the technique. The results from the tests are recorded. The performance evaluation metrics used are accuracy and f1-score even though precision, recall, runtime, and prediction latency are also shown. The pseudocode representing the hybrid technique for this system is shown below.

```
Input:
B.O.W vectors set X;
Stacked ensemble F:
Base estimator model f;
Final estimator model g;
F = f + g #stacked ensemble
Process F(X):
base_preds = f(x);
final_preds = cross_validate(g(base_preds))
return final_preds
Output:
final_preds.
```

3. RESULT AND ANALYSIS

This chapter presents a concise form of the results from our experiments categorized by research question.

RQ1: How would hybrid machine learning techniques for performing comparative opinion mining be evaluated?

To answer this question, we experimented with different hybrid machine learning techniques on dissimilar datasets and domains. We then compared the performances of our techniques. Our study established that accuracy is a good measure of algorithm performance if the data has balanced classes. However, if the classes are not well balanced in the dataset, then the f1-score should be used as it represents the best combination of precision and recall [38]. To identify the best-performing hybrid machine learning algorithm carrying out comparative opinion mining on online customer reviews, we tabulated the results for each hybrid technique. Based on the accuracy and f1-score metrics, we were able to identify the best-performing hybrid machine learning technique for comparative opinion mining. After we completed the evaluation of the performance of the various techniques, this study recommends the MLP + RF hybrid technique because it outperformed the other techniques in both accuracy and f1-score.

RQ2: What is the efficiency of different hybrid machine learning techniques in comparative opinion mining?

To find a solution to this research query, we implemented different hybrid machine learning algorithms on three datasets to perform comparative opinion mining. We then evaluated the performance of these techniques using accuracy and f1-score. The experiments are detailed below. All the hybrid techniques had MLP as the base estimator with each of these four single classifiers as the final estimator: DT, RF, SGD, and SVM. Our results include other evaluation measures such as precision, recall, runtime, and prediction latency are presented as well.

Experiment #1:

In this experiment, we applied independent traditional machine learning techniques and one deep learning technique to perform comparative opinion mining on online customer reviews. In particular, the deep learning technique we worked with was the Multi-Layer Perceptron (MNP) largely because it uses the same vectorization approach as the traditional machine learning techniques. From this experiment, we wanted to determine the most optimal independent machine learning classifiers that could be used to build our hybrid machine learning techniques.

Table 3. Performance of the Independent Machine Learning Techniques Across the Three Datasets

Classifier	Accuracy (%)			F1-Score (%)			Averages (%)		Efficiency			
ML Technique	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 Pearl Int'l vs Marriott)	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 (Pearl Int'l vs Marriott)	Accuracy	F1-Score	D1	D2	D3	Avg
DT	83.8	91.8	76.0	83.9	91.8	76.8	83.9	84.2	0.0	0.0	0.0	0.0
KNN	47.1	53.7	55.9	57.9	59.8	61.5	52.2	59.7	0.0	0.0	0.0	0.0
LR	86.0	92.0	78.2	86.1	92.0	79.1	85.4	85.7	0.0	0.0	0.0	0.0
MLP	*86.2*	*92.6*	*77.9*	*86.4*	*92.6*	*78.7*	*85.6*	*85.9*	*0.1*	*0.1*	*0.0*	*0.1*
MNB	80.4	88.8	73.5	81.1	88.9	75.1	80.9	81.7	0.0	0.0	0.0	0.0
RF	85.3	92.3	76.7	85.5	92.4	77.9	84.8	85.3	0.0	0.0	0.0	0.0
SGD	*86.6*	*91.8*	*78.6*	*86.7*	*91.8*	*79.3*	*85.7*	*86.0*	*0.0*	*0.0*	*0.0*	*0.0*
SVM	76.5	86.3	74.1	78.1	86.5	75.8	79.0	80.2	0.1	0.1	0.0	0.1

From the above statistics, it is evident that the SGD technique on average outperformed all the other single classification techniques in average accuracy (85.7%) as well as average f1-score (86.0%). The Multi-Layer Perceptron (MLP) technique performed very well, too, with an insignificant variation (0.1%) from the performance of the SGD technique. MLP's average accuracy is 85.6% while its f1-score is 85.9%. For this reason, each of these two techniques would be suitable as an estimator in the hybrid architecture for comparative opinion mining. However, we needed a deep learning technique as the base estimator because of the need to handle larger datasets where deep learning techniques often perform better than traditional machine learning techniques. This is why we chose MLP as our base estimator. This stacking approach was similarly used in a study by [28] to create a hybrid technique consisting of LSTM and CNN, both being deep learning techniques. This experiment was done to given us an overview of how the single machine learning techniques perform so we could select the best for hybridization.

Experiment #2:

We carried out this experiment on dataset 1, which contained 3000 online reviews about "Microsoft vs Google". Table 4 below displays the results of this experiment. The results show that both MLP + DT hybrid techniques outperformed the other hybrid techniques in both accuracy and f1-score. It had an average accuracy of 86.6% and an f1-score of 86.8%. It is evident from these results that four hybrid classifiers performed satisfactorily, having a minimum accuracy of 85.4% and a minimum f1-score of 85.6%.

Table 4. Performance of Hybrid Machine Learning Techniques On Dataset #1

Hybrid ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	86.0	86.7	86.0	86.1	59.7	0.0
MLP + RF	86.6	87.8	86.6	86.8	77.7	0.1
MLP + SGD	86.9	87.8	86.9	87.1	61.1	0.0
MLP + SVM	85.6	86.9	85.6	85.8	60.7	0.2

Experiment #3:

We carried out this experiment on dataset 2, which contained 3000 online reviews about "Facebook vs Twitter". Table 5 below shows that the MLP + RF and MLP + SGD hybrid techniques outperformed the other hybrid classifiers in both accuracy and f1-score. They both had an accuracy of 92.4% and an f1-score of 92.5%. Generally, all classifiers performed well in on this dataset, yielding a minimum accuracy of 91.6% and f1-score of 91.6%.

Table 5. Performance of the Hybrid Machine Learning Techniques On Dataset #2

Hybrid ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	92.0	92.4	92.0	92.0	52.2	0.0
MLP + RF	92.6	92.9	92.6	92.6	63.4	0.1
MLP + SGD	92.3	92.7	92.3	92.3	50.2	0.0
MLP + SVM	91.9	92.6	91.9	91.9	57.8	0.2

Experiment #4:

We carried out this experiment on dataset 3, which contained 1000 online reviews about "Pearl Continental vs Marriott". Table 6 below shows the results from this experiment as regards the application of different hybrid machine learning classifiers on this dataset. The results show that two of the four hybrid models achieved 100% performance in both accuracy and f1-score. This is probably because of overfitting due to the relatively smaller size of this dataset. The MLP + DT and MLP + RF hybrid classification techniques had accuracies of 99.0% and 99.7%, respectively. These accuracies we probably affected by the

overfitting of the model on the dataset. However, comparing with the 93.8% accuracy obtained in dataset 1 and 2 shows a variance of 5.2%. This variance is not significant enough to affect the reliability of the classifier in terms of predictive accuracy.

Table 6. Performance of the Hybrid Machine Learning Techniques On Dataset #3

Hybrid ML Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Runtime (ms)	Prediction Latency (ms)
MLP + DT	98.4	98.4	98.4	98.3	5.5	0.0
MLP + RF	99.7	99.7	99.7	99.7	5.9	0.0
MLP + SGD	91.8	93.7	91.8	91.6	5.1	0.0
MLP + SVM	99.7	99.7	99.7	99.7	5.4	0.0

RQ3: Which one is the most efficient hybrid machine learning technique for comparative opinion mining?

To answer this research question, we tested different single machine learning algorithms and one deep learning algorithm on the same datasets and features to perform comparative opinion mining. We then evaluated the performance of the classifiers using accuracy, f1-score, runtime, and prediction latency. Iterative experiments were conducted by changing the datasets for all classification algorithms. Our findings show that the MLP + RF and MLP + SVM hybrid machine learning techniques produced the best result in both model training time (runtime) and prediction latency. All the hybrid machine learning techniques we implemented on the three datasets performed satisfactorily on all measures including model training runtime and prediction latency. The efficiency of these two techniques varied from the most accurate hybrid technique by 0.8 milliseconds, which is insignificant).

Table 6. Performance Summary of the Hybrid Machine Learning Techniques On Datasets 1-3

Classifier	Accuracy (%)			F1-Score (%)			Averages (%)		Efficiency			
Hybrid ML Technique	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 (Pearl vs Marriott)	D1 (Microsoft vs Google)	D2 (Facebook vs Twitter)	D3 (Pearl vs Marriott)	Accuracy	F1-Score	D1	D2	D3	Avg
MLP + DT	86.0	92.0	98.4	86.1	92.0	98.3	92.1	92.1	0.0	0.0	0.0	0.0
MLP + RF	*86.6*	*92.6*	*99.7*	*86.8*	*92.6*	*99.7*	*93.0*	*93.0*	*0.1*	*0.1*	*0.0*	*0.1*
MLP + SGD	86.9	92.3	91.8	87.1	92.3	91.6	90.3	90.3	0.0	0.0	0.0	0.0
MLP + SVM	85.6	91.9	99.7	85.8	91.9	99.7	92.4	92.5	0.2	0.2	0.0	0.1

The above results show that the MLP + RF hybrid classifier outperformed other classifiers, across the three datasets. The accuracy was 93.0% and the f1-score was 93.0%. Generally, the hybrid techniques had a minimum average accuracy and f1-score of 90.3%, implying that the techniques are reliable for use in comparative opinion mining.

Q4: How suitable would a hybrid machine learning technique be in comparative online reviews classification?

To answer this question, we considered the most critical algorithm performance metrics in carrying out classification tasks. From our findings, we established that both accuracy and f1-score are important metrics in evaluating the performance of machine learning techniques. Accuracy is preferably used if the classes are well-balanced while F1-score is more preferred where classes are imbalanced as it represents a good balance between precision and recall. The lowest accuracy achieved was 90.3% and the highest accuracy was 93.0%. The lowest f1-score was 90.3% while the highest f1-score was 93.0%. This makes our developed hybrid machine learning models a more reliable choice compared to using human resources in classifying comparative opinion reviews. This reason is besides the fact that computer models perform the classification tasks with a speed that humans cannot match. For instance, in terms of time efficiency, the least performing hybrid machine learning model had an average latency of 0.1 milliseconds, which cannot be achieved by human classifiers. The prediction latency is a measure of how long the model took, in milliseconds, to output the classification results, given a dataset. Therefore, hybrid machine learning classification techniques are highly suitable for carrying out comparative opinion mining.

Discussion

From the results above, it is evident that the best-performing hybrid machine learning technique for comparative opinion mining is the MLP + RF with an average accuracy of 93.0% and f1-score of 93.0%. These measures were computed across the three datasets used. The hybrid techniques we developed had a minimum average accuracy and f-1-score of 90.3%, meaning, they could all be reliably applied to comparative opinion mining. The variance between these averages and those of the best hybrid classifier is

0.1% for accuracy and f1-score, and 0.1 milliseconds for prediction latency. Therefore, these hybrid machine learning algorithms have satisfactory performance hence could be reliably applied to comparative opinion mining for applications like brand reputation monitoring [37]. We observed that our hybrid techniques outperformed all single machine learning techniques by a significant margin of 7.3% in accuracy and 7.0% in f1-score. This variance was obtained from the difference between the best single technique (SGD) and the best hybrid machine learning technique (MLP + RF). Further to this, MLP had an accuracy and f1-score of 85.6% and 85.9% when used alone. RF had an accuracy and f1-score of 84.8% and 85.3% respectively. The average accuracy and f1-score of these two independent machine learning techniques is 85.2% and 85.6% respectively. On the other hand, our hybrid machine learning technique for comparative opinion mining that fused together MLP and RF attained a significantly higher performance with an accuracy of 93.0% and f1-score of 93.0%. This is a performance gain of 7.8% and 7.4% in accuracy and f1-score.

4. CONCLUSION

Throughout this research, our focus was to develop and empirically evaluate the performance of different hybrid machine learning techniques upon application to comparative opinion mining of comparative opinionated data. To do this, we followed a five-step process: (1) Collection of datasets for comparative opinion mining, (2) pre-processing of the collected data, (3) development of hybrid machine learning techniques, (4) applying the different hybrid machine learning techniques to perform comparative opinion mining on comparative online reviews, (5) comparing the performance results of the different machine learning algorithms using accuracy and f1 score metrics and (6) recommending the best-performing machine learning classifier.

Our proposed approach aimed at determining the various sentiment polarity classes from classifying comparative opinion texts. In this study, the hybrid machine learning techniques that were applied to comparative opinion mining were: MLP + DT, MLP + RF, MLP + SGD, and MLP + SVM. To evaluate the performance of these hybrid machine learning algorithms, accuracy, and f1-score performance metrics were utilized in addition to prediction latency to help determine time-based efficiency. From our observations, the best-performing hybrid algorithm across most of the datasets was the MLP + RF. Generally, all of the hybrid techniques implemented in our experiments had a minimum of 93.0% in both accuracy and f1-score measures across the three datasets.

Limitations

This study was centered on three application domains: (1) technology, (2) social media, and (3) hospitality. This increased the reliability of the models/techniques used when compared with previous studies that relied on datasets belonging to the same domain or just one dataset. Further to this, the application of the random split method in generating the training set and testing set meant that there was reduced overfitting of data. This study faced a few limitations. (1) However, the results obtained may not be robust enough because of the number and size of datasets used, (2) the imbalanced nature of the comparative datasets used may have degraded the performance of the hybrid techniques, (3) further, the performance of the machine learning algorithms degraded as we included three sentiment classes per entity [39] when compared to studies that only used two classes per entity.

Future Directions

This study recommends the following for future studies: (1) the use of even larger datasets to potentially attain more reliable results, (2) a deliberate effort to use balanced datasets to improve classifier performance, (3) experimenting purely with comparative datasets, (4) the use of alternative dataset splitting methods like cross-validation and stratified sampling in the evaluation of how the various machine learning algorithms perform.

REFERENCES

[1] B. Liu, *Sentiment analysis and opinion mining*. Cham, Switzerland: Springer, 2012.
[2] K. D. Varathan, A. Giachanou, and F. Crestani, "Comparative opinion mining: A review," *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 811–829, Apr. 2017, doi: 10.1002/asi.23716.
[3] M. M. Eldefrawi, D. S. Elzanfaly, M. S. Farhan, and A. S. Eldin, "Sentiment analysis of Arabic comparative opinions," *SN Appl. Sci.*, vol. 1, no. 5, p. 411, May 2019, doi: 10.1007/s42452-019-0402-y.
[4] Y. Li, B. Jia, Y. Guo, and X. Chen, "Mining User Reviews for Mobile App Comparisons," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–15, Sep. 2017, doi: 10.1145/3130935.

[5] S. Li *et al.*, "Product comparison using comparative relations," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, Beijing China: ACM, Jul. 2011, pp. 1151–1152. doi: 10.1145/2009916.2010094.
[6] K. Xu, S. S. Liao, J. Li, and Y. Song, "Mining comparative opinions from customer reviews for Competitive Intelligence," *Decision Support Systems*, vol. 50, no. 4, pp. 743–754, Mar. 2011, doi: 10.1016/j.dss.2010.08.021.
[7] R. Feldman, M. Fresko, J. Goldenberg, O. Netzer, and L. Ungar, "Extracting Product Comparisons from Discussion Boards," in *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, Omaha, NE: IEEE, Oct. 2007, pp. 469–474. doi: 10.1109/ICDM.2007.27.
[8] T. Kurashima, K. Bessho, H. Toda, T. Uchiyama, and R. Kataoka, "Ranking Entities Using Comparative Relations," in *Database and Expert Systems Applications*, S. S. Bhowmick, J. Küng, and R. Wagner, Eds., in Lecture Notes in Computer Science, vol. 5181. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 124–133. doi: 10.1007/978-3-540-85654-2_15.
[9] J. Sun, C. Long, X. Zhu, and M. Huang, "Mining Reviews for Product Comparison and Recommendation," *Polibits*, vol. 39, pp. 33–40, Jun. 2009, doi: 10.17562/PB-39-5.
[10] U. Younis, M. Z. Asghar, A. Khan, A. Khan, J. Iqbal, and N. Jillani, "Applying Machine Learning Techniques for Performing Comparative Opinion Mining," *Open Computer Science*, vol. 10, no. 1, pp. 461–477, Dec. 2020, doi: 10.1515/comp-2020-0148.
[11] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *FNT in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008, doi: 10.1561/1500000011.
[12] M. Sun, X. Huang, H. Ji, Z. Liu, and Y. Liu, Eds., *Chinese computational linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18-20, 2019: proceedings*. in Lecture notes in computer science Lecture notes in artificial intelligence, no. 11856. Cham, Switzerland: Springer, 2019.
[13] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis without aspect keyword supervision," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, San Diego California USA: ACM, Aug. 2011, pp. 618–626. doi: 10.1145/2020408.2020505.
[14] A. U. R. Khan, M. Khan, and M. B. Khan, "Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining," *Procedia Computer Science*, vol. 82, pp. 57–64, 2016, doi: 10.1016/j.procs.2016.04.009.
[15] S. Banihashemi, G. Ding, and J. Wang, "Developing a Hybrid Model of Prediction and Classification Algorithms for Building Energy Consumption," *Energy Procedia*, vol. 110, pp. 371–376, Mar. 2017, doi: 10.1016/j.egypro.2017.03.155.
[16] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artif Intell Rev*, vol. 54, no. 7, pp. 4997–5053, Oct. 2021, doi: 10.1007/s10462-021-09973-3.
[17] S. Yildirim, "Comparing Deep Neural Networks to Traditional Models for Sentiment Analysis in Turkish Language," in *Deep Learning-Based Approaches for Sentiment Analysis*, B. Agarwal, R. Nayak, N. Mittal, and S. Patnaik, Eds., in Algorithms for Intelligent Systems. Singapore: Springer Singapore, 2020, pp. 311–319. doi: 10.1007/978-981-15-1216-2_12.
[18] H. Lu, L. Ehwerhemuepha, and C. Rakovski, "A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance," *BMC Med Res Methodol*, vol. 22, no. 1, p. 181, Dec. 2022, doi: 10.1186/s12874-022-01665-y.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[20] A. Addiga and S. Bagui, "Sentiment Analysis on Twitter Data Using Term Frequency-Inverse Document Frequency," *JCC*, vol. 10, no. 08, pp. 117–128, 2022, doi: 10.4236/jcc.2022.108008.
[21] J. Yan, "Text Mining with R: A Tidy Approach, by Julia Silge and David Robinson. Sebastopol, CA: O'Reilly Media, 2017. ISBN 978-1-491-98165-8. XI + 184 pages.," *Nat. Lang. Eng.*, vol. 28, no. 1, pp. 137–139, Jan. 2022, doi: 10.1017/S1351324920000649.
[22] J. Jin, P. Ji, and R. Gu, "Identifying comparative customer requirements from product online reviews for competitor analysis," *Engineering Applications of Artificial Intelligence*, vol. 49, pp. 61–73, Mar. 2016, doi: 10.1016/j.engappai.2015.12.005.
[23] S. Rana and A. Singh, "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," in *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, India: IEEE, Oct. 2016, pp. 106–111. doi: 10.1109/NGCT.2016.7877399.

[24] H. M. Ismail, S. Harous, and B. Belkhouche, "A Comparative Analysis of Machine Learning Classifiers for Twitter Sentiment Analysis," *RCS*, vol. 110, no. 1, pp. 71–83, Dec. 2016, doi: 10.13053/rcs-110-1-6.
[25] R. Joshi and R. Tekchandani, "Comparative analysis of Twitter data using supervised classifiers," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India: IEEE, Aug. 2016, pp. 1–6. doi: 10.1109/INVENTIVE.2016.7830089.
[26] A. Ejaz, Z. Turabee, M. Rahim, and S. Khoja, "Opinion mining approaches on Amazon product reviews: A comparative study," in *2017 International Conference on Information and Communication Technologies (ICICT)*, Karachi: IEEE, Dec. 2017, pp. 173–179. doi: 10.1109/ICICT.2017.8320185.
[27] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining Knowl Discov*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1253.
[28] K. Abu Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, "LSTM-CNN Deep Learning Model for Sentiment Analysis of Dialectal Arabic," in *Arabic Language Processing: From Theory to Practice*, K. Smaïli, Ed., in Communications in Computer and Information Science, vol. 1108. Cham: Springer International Publishing, 2019, pp. 108–121. doi: 10.1007/978-3-030-32959-4_8.
[29] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: Trends and challenges," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, Doha, Qatar: IEEE, Nov. 2014, pp. 148–155. doi: 10.1109/AICCSA.2014.7073192.
[30] M. Tkachenko and H. W. Lauw, "Generative Modeling of Entity Comparisons in Text," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai China: ACM, Nov. 2014, pp. 859–868. doi: 10.1145/2661829.2662016.
[31] X. Fan, C.-H. Lung, and S. A. Ajila, "Using Hybrid and Diversity-Based Adaptive Ensemble Method for Binary Classification," *IJIS*, vol. 08, no. 03, pp. 43–74, 2018, doi: 10.4236/ijis.2018.83003.
[32] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz, "Fair train-test split in machine learning: Mitigating spatial autocorrelation for improved prediction accuracy," *Journal of Petroleum Science and Engineering*, vol. 209, p. 109885, Feb. 2022, doi: 10.1016/j.petrol.2021.109885.
[33] D. Effrosynidis, G. Peikos, S. Symeonidis, and A. Arampatzis, "DUTH at SemEval-2018 Task 2: Emoji Prediction in Tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 466–469. doi: 10.18653/v1/S18-1074.
[34] N. I. Khairi, A. Mohamed, and N. N. Yusof, "Feature Selection Methods in Sentiment Analysis: A Review," in *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*, Marrakech Morocco: ACM, Mar. 2020, pp. 1–7. doi: 10.1145/3386723.3387840.
[35] A. Sharifi and K. Alizadeh, "A Novel Classification Method Based on Multilayer Perceptron-Artificial Neural Network Technique for Diagnosis of Chronic Kidney Disease," *Ann Mil Health Sci Res*, vol. 18, no. 1, May 2020, doi: 10.5812/amh.101585.
[36] S. Diab, "Optimizing Stochastic Gradient Descent in Text Classification Based on Fine-Tuning Hyper-Parameters Approach. A Case Study on Automatic Classification of Global Terrorist Attacks," 2019, doi: 10.48550/ARXIV.1902.06542.
[37] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif Intell Rev*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
[38] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognition Letters*, vol. 30, no. 1, pp. 27–38, Jan. 2009, doi: 10.1016/j.patrec.2008.08.010.
[39] D. Silva-Palacios, C. Ferri, and M. J. Ramírez-Quintana, "Improving Performance of Multiclass Classification by Inducing Class Hierarchies," *Procedia Computer Science*, vol. 108, pp. 1692–1701, 2017, doi: 10.1016/j.procs.2017.05.218.

BIBLIOGRAPHY OF AUTHORS

Bernard Omoi Ondara is a Lecturer at the Department of Computing and Information Technology, Kenyatta University. He is currently pursuing a PhD in Computer Science at Kenyatta University. He completed his MSc. in Computer Science from the University of Nairobi and his B.Sc. in Information Sciences – IT Major from Moi University. He has interest in Artificial Intelligence, Machine Learning, Data Science, Big Data Technologies, and Digital Marketing.

Stephen T. Waithaka is an Executive Dean at the Digital School of Virtual and Open Learning. He is a former Senior Lecturer, in the Department of Computing and Information Technology, Kenyatta University. Dr. Waithaka has a PhD in Information Systems from Kenyatta University. He completed his MSc. In Information Technology and BSc. in Information Technology from EBS, UK. He is also a Member of the editorial board of Horizon Research Publishing (HRPUB) journal of Computer Science and Information Technology and Engineering International (EI) journal. His special interests are in Artificial Intelligence, Applied Information & Communication Techonlogies, and Computer Networks.

John M. Kandiri is a senior lecturer in the Department of Computing and Information Technology, Kenyatta University. He has a PhD in Information System, BSc. in Information Systems and BSc. in Information Sciences – IT Major. Dr. Kandiri is also a Microsoft Certified Developer (MCP). He has a Certificate in Applied Research from Steadman, and another ceriticate in Case Writing. He is a curriculum developer with African Virtual University (AVU). His special interests are in Artificial Intelligence and Data Structures and Algorithms.

Lawrence Muchemi is a senior lecturer in the Department of Computing and Informatics at The University of Nairobi, Kenya. He completed his Ph.D in Computer Science from the Universty of Nairobi and his B.Technology Engineering and M.Phil. Engieering-AI Systems from Moi Univeristy.
His special interests are in Artificial Intelligence and Machine Learning with special focus on Natural Language Processing.