ISSN 2347 - 3983

Volume 8. No. 9, September 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter237892020.pdf

https://doi.org/10.30534/ijeter/2020/237892020

Feature-Based Sentimental Analysis On Product Review System Using CUDA-BB Algorithm

Siva Kumar Pathuri[1], Dr.Anbazhagan.N[2]

[1]Koneru Lakshmaiah Education Foundation, India, spathuri@kluniversity.in

[2]Alagappa University, India, anbazhagann@alagappauniversity.ac.in

ABSTRACT

Sentiment analysis methods are classified as feature extraction method, i.e., view's, review's or sentence to predict the emotion of a sentence or a text using natural languages processing(NLP). The analysis of sentiment involves classification of the text into three phases "positive," "negative" or "The Neutral." It Analyzes the data and labels and classified as either "good" "better" "best" or "bad", "worse" based on the emotions or feeling given by customer and finally classified as positive and negative or neutral respectively. So, in regard for the past few years, The World Wide Web (WWW) has become a vast source in providing raw data which is is in the form of opinions or emotions or reviews given by the user or customer about a particular product. E -commerce, has grabbed the attention of the business people to improve the quality of their product by taking the review's from different Social media websites like Facebook , Twitter, Amazon, Flipkart, etc Sentimental analysis or Opinion mining is one of the major challenges of NLP (natural language processing).Business Analytics plays a very important role in the current scenario. In particular, these people rely on the feedback of their products given by the customer's to withstand the competition and knowledge mining that can give them an outstanding view of what to expect in the future..Classification and rule induction describes the main topics in the field of decision making and knowledge discovery. In this paper, we propose CUDABB (CUDA Bag-Boost) algorithm is used to find the overall star rating of Smartphone's using GPU parallel computing. This strategy not only reduces variance and bias but also this approach allows to produce better predictive performance compared to a single model. And the results obtained is compared with SLIQ and MMDBM using pycuda and GPU with computed acceleration rate(speedup) time using Amazon mobile review dataset. The aim of GPU mining technique is to enhance the execution speed with less handling time. Finally, we conclude that the proposed method achieves better accuracy.

Key words: Classification Techniques, CUDABB, GPU, Mobile Reviews

1. INTRODUCTION

The way data is extracted from a huge number of data sets is described by data mining. The data mining approaches include sorting, aggregation rules, clusters and other techniques. One of the most important data mining tools is classification. In classification we provided a variety of sample records, in which each record consists of many attributes, called a training data set. A number or segment [1,2] can be an attribute. The attribute is referred to as a numerical attribute (for instance, cost , camera pixel etc), where the attributes fall within the ordered domain values. According to the latest study by Global Digital Forensics', the world rules on online social media, tablets, mobiles, computers and many more Smart phones etc for gathering the review's given by the product consumers will continue to be at the forefront of business and personal use in the near future to withstand in the market which is the key objective for us. A powerful broad data processing and analysis provides intelligent and useful knowledge. With the increase of Internet connectivity growth, we are able to evaluate massive volumes of data day to day and to predict consumer's needs and future demands. E-commerce has been exploding around the world in recent years and most people prefer to buy products from these websites. Therefore, the form of reviews involves large amounts of data Produced to assist potential buyers in selecting the right product. These reviews also contain opinions that could be helpful for the company to identify the areas to be improved. However, every review on the product is impractical for the user. In addition, only a few reviews may give a partial idea of the product. Some reviews may not have credible sources, which the users can't differentiate. In this paper the customer's reviews are taken as input, i.e. specifically, concentrated on positive and negative reviews. On the basis of the tests, we will estimate the effectiveness of the review and the rating depending on the review content[3]. For example, "This mobile is excellent, but cost is too high" used the emotional trend between the views of mobile users as part of the speech label. This is done in two steps the first step is Aspect Extraction, and the second phase is polarity. o perform classification techniques, the comments and rankings based on the text review were used to find the effectiveness of the product. Classification can be used as step in the development of knowledge in data mining. We need all to be done in time in the field of internet and fast-growing technology. This paper introduces a

Hybrid classification technique in classification mining called CUDABB Algorithm (Bag Boost). Using this Hybrid approach, we can reduce the time taken for sorting and to classify the data[4].

In this paper, we summarize the key contributions:

1) First, predictive analysis was used for Smartphone review ratings. Second, average Smartphone reviews far lesser than PC. The third set of reviews' length and count is based on three distinct polarities – positive, negative, and neutral. Such statistical characteristics are essential for designing and classification in the evaluation process.

2) A number of comparative experiments have been carried out in order to find more appropriate methods for commenting on short texts[5]. These test sets include (1) Comparison of polarity classification algorithms (2) the comparison of text representation methods. (3) We divided the data into the different groups by number of words in order to have an impact on short text or on long text. These tests help us find a more effective and precise method.

3) The above-mentioned experiment is built on a large, real-world data set. i.e, over 5,000,000 + real mobile reviews are taken from Kaggle repository. The key objective of this paper is to predict star rating of a mobile phone by using different classification techniques so, this paper moves the conventional computer data processing to mobile devices with some rational improvements to trace customer behavior, regarding reviews on products and protect against fraud gossips form different competitors[6]. So from this the mobile vendors can take best decision for developing their future products. It can be done with the help of a 2 step method for predicting star rating of a mobile 1) Data Preprocessing 2) Feature Extraction and Polarity classification 3) CUDABB Algorithm to find the accuracy of the model[7].

We have implemented an effective CUDABB Classifier on GPU which uses various threads. And the time taken by each CUDA thread is used to find the results of each individual CUDA thread. Generally in GPU mining, the code for finding the mid-point is given as input to CUDABB algorithm, which in turn used to compute acceleration ratio (speed-up) time, i.e., used to calculate the class count values. In GPU the data is divided into small threads and calculating class count values of each and every thread tell that GPU computing is fastest when compared with CPU computing. In this paper there are 6 sections ie section 2 tells about related work. Section 3 shows the proposed method CUDABB algorithm. Section 4 shows the experimental results and comparison with existing 2 algorithms. Finally in Section 5 tells about Conclusion and Future Work[8].

2. RELATED WORK

2.1 Product Review

Subjectivity detection and feeling prediction, aspect based sentiment overview, and Text Opinion summarization, Feature extraction of product and spam detection. are the main areas of research in sentiment analysis. The detection of subjectivity is a task to see whether or not the text is conveyed. The prediction of feelings concerns the prediction of the text polarity whether positive or negative. The summary of feelings generates sentiment in form of star ratings or numerical characteristics of the product. For example, let us take one product review given in social websites i.e., amazon.com:

I feel so LUCKY to have found this used (phone to us & not used hard at all), phone online from someone who upgraded and sold this one. My Son liked his old one that finally fell apart after 2.5+ years and didn't want an upgrade!! Thank you Seller, we really appreciate it & your honesty.

After analyzing the text with the help of text mining based on the classification review of the product a decision rule is obtained as follows:

Lucky->Positive, Liked->Positive, Appreciate->Positive.

2.2 Sentimental Analysis

Sentiment classification is carried out at three levels Text level, Sentence level and Aspect level or attribute level. The role at document level is to classify the records into 3 types of polarity ie positive, negative or neutral level. Sentence level classification classifies the sentence as a positive, negative, neutral class depending on the level of each sentence. Firstly, the polarity of each sentence is determined and then the total meaning of the sentence is measured. Aspect or Feature level Sentiment Classification defines and extracts the product Sentiment analysis[9]. This process can be divided into four steps: text selection, pre-processing, Transformation and Feature Extraction[10].

1) Describe a data set area, such as data sets, product reviews, reviews goods, tweets, and others, which spanned a region.

2) Pre-processing: first processing step, usually done by tokenization process, elimination of stop words and stemming process[11].

3) Transformation: representational figures from textual data measured method. Binary representation widely and specifically used count the document's appearance or lack of a word. How many times a word appears as a weighting scheme is often used as semantic data.

4) Feature Selection: Feature Collection: The selection of features (feature selection) will make the classifier more efficient / effective by minimizing the amount of data to be processed in order to determine the appropriate features for further analysis[12].

In this paper we compare the accuracy of main classification based algorithms such as SLIQ, MMDBM. and CUDABB algorithm. SLIQ (Supervised Learning In Quest) is a fast scalable tree classifier in data mining, which can handle both numerical and categorical data for large attributes, It uses presorting technique during tree construction phase which inturn reduces the cost of evaluating the attributes and results in a compact and accurate trees. This sorting technique combines with a breadth-first tree growing strategy to enable disk-resident datasets classification. SLIQ uses a MDL principle for tree-pruning [1]. Mixed Mode Database Miner (MMDBM) is a one of the new tree classifier which can handle both numerical and categorical data. It is a 2 step algorithm ie the 1st one is a predictive classifier which gives description of the algorithm and 2nd one is object oriented implementation [3,4].CUDA Bag-Boost Algorithm (CUDABB) is a new decision tree classifier which can handle both numerical and categorical attributes in large datasets. It is a 2 step process in which the 1st step aspect extraction and 2nd step Polarity classification based on the above 2 steps the accuracy of the classifier is calculated.

2.3 Feature Selection

In order to select a feature of a product some statistical analysis on a product is to be considered[13].

1) Seeing the relation between reviews and rating.
2) Seeing the connection between feedback and ratings.
3) Consider the relationship between price and ranking.
4) Determining the ranking of the brands.
5) Noting the count of word repeatedly used by the consumer.
6) Figuring out the polarity between each and every review.

Example: A person is trying to buy a mobile phone in online or buying some product he has to take some opinion’s or review’s which are given by the product user’s which will be more helpful for them to take a decision. With the help of the decisions taken by the customers it is possible to make one's business success [23,24].

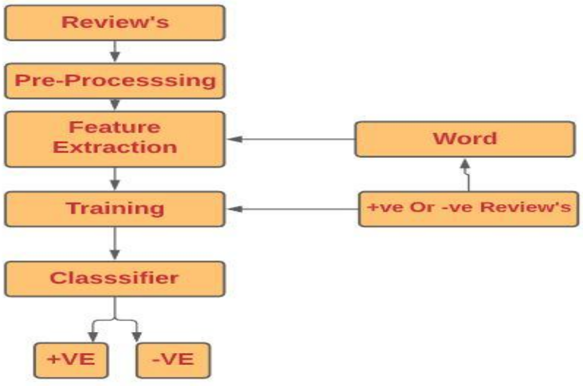

Figure 1: System Architecture

The figure.1 shows about how the system components as such as Feature extraction from different social media, preprocessing of data, feature extraction, the Training set is for the given finding the accuracy.

2.4 Polarity Classification

The training set is obtained by a predefined set in which polarity can be determined using Vader’s Sentimental analysis obtained as positive, negative, or neutral. The Classifier classifies the reviews according to the training set and regulates the polarity of each review as the output. We test the way of generating such data using emoticons (e.g. # best mood, # Wonderful I # love) to classify favourable, negative , and neutral ratings to be used for classifier training[25]. These ratings and hashtags are also important to evaluate how people think about a product. VADER sentiment analyzer is a model used for text sentiment analysis that is adaptive to calculate both polarity (positive / negative) and frequency (strength) of emotions.

For Eg:

a = 'This was a Worst Mobile ever.”By applying Vader’s sentimental analysis on the above text with the help of predefine method polarity_scores(a) the text is classified as{'negative': 0.9, 'neutral': 0.1, 'positive': 0, 'compound': 0.8404}

Similarly a = 'This was the best and awesome Smartphone never before ever after!!!'.By applying Vader’s[14] sentimental analysis on the above text with the help of predefine method polarity_scores(a) the text is classified as{'negative': 0.0, 'neutral': 0.425, 'positive': 0.575, 'compound': 0.8877}

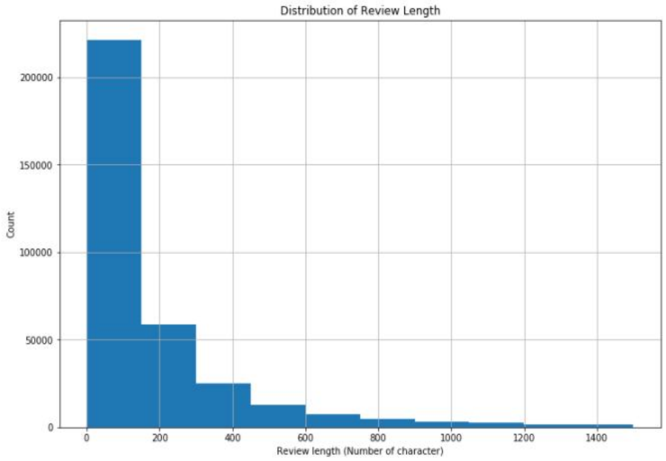

Figure 2: Gives the count vs review length.

3. METHODOLOGY

The work is mainly based on 3 classification algorithms such as SLIQ, MMDBM and CUDABB. SLIQ(Supervised Learning In Quest) is one of the fastest classifier in data mining. SLIQ is decision tree classification technique which can handle both numerical and categorical data[15,16] and also to classify large training data. It is used to build compact and accurate trees and also works on the principle of pre-sorting technique in the tree-growth phase to reduce the cost of evaluating numeric attributes [17].

Mixed Mode Database Miner (MMDBM) is yet another decision tree classifier which can handle both numerical and categorical attributes in large datasets. The algorithm is divided into two parts. The first one is predictive classifier that gives detailed description of the algorithm and second one is object oriented design, that gives the object oriented implementation[18,19].

Machine learning plays a major area in the field of AI&DS which uses some statistical data for analyzing and tells the machine what to do.Fig3 shows the procedure for predicting the o/p of the designed model.

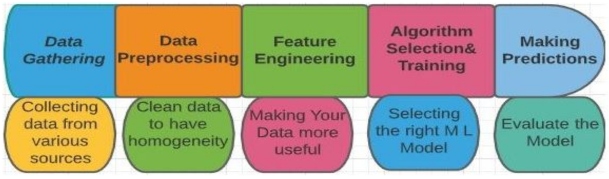
Data Gathering
Data Preprocessing
Feature Engineering
Algorithm Selection& Training
Making Predictions
Collecting data from various sources
Clean data to have homogeneity
Making Your Data more useful
Selecting the right M L Model
Evaluate the Model

Figure 3: Procedure for predicting O/P of the Designed Model

They are three types of classification techniques in machine learning:

1) Supervised: This creates a model for known input and output data so that it can predict future outputs or labeled data.

2) Un-Supervised: This finds hidden patterns or intrinsic structures in input data or un-labeled data.

3) Reinforcement Learning: This uses a concept called Trail and Error method.

A. Supervised Learning

In this type of learning method it uses the labeled data. With the known inputs and outputs it predicts. So, that it can predict the future outputs. Supervised algorithms are Naïve-Bayes, Random Forest, Decision Tree, Xg Booster, etc.

B.Un- Supervised Learning

In this type of learning method it uses the unlabelled data. It will find the hidden patterns. The unsupervised algorithms are Hierarchical clustering, GaussianMixturemodels,Hidden Marko models, Self-Organizing maps, Fuzzy-means clustering, subtractive learning and k-means clustering[20,21] etc.

C. Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning which deals with trial and error method. And it is about making decisions step by step. In other war RL is defined as the ouput depends on the present input as well as next input. Since the decisions are dependent so we have to assign lables in sequence for each independent decision taken. Eg: Chess Game[22]

3.1. Procedure

During this classification process, the data is split into two sections. Training and testing, i.e. The training set train's the classifier and the test set gives the classification performance. In general several methods of machine learning which classify emotions or reviews given by several customers. The data set includes the opinions of a product provided by various consumers. Dataset is taken from the Kaggle repository named as Amazon Unlocked Mobile Reviews. The product consumer shares his / her own opinion on his / her perceptions and gives a rating on a scale of 1 to 5. The overall value is based on all consumer ratings for the final product categorization. Additional reviews on the product will also be useful based on the functionality, thereby adding value to the review and the reviewers in turn also. In this article, we explored the relation between price, reviews and ratings over 500000 + opinions on smart phones. Datasets are obtained from:/www.kaggle.com.The data set contains the following information or attributes from the 1. Product Title 2.Brand Name 3.Price 4.Rating 5. Text review 6. Review Votes. And the data taken is of the form.csv format. Such details was used to determine what the customer's ranking would have offered, based on the emotions contained in their feedback. The data collection is allocated to training data and evaluation data for 260000 + records and 2,40152 + records respectively for performing classification. Figure 7 shows about how to train and test the data.

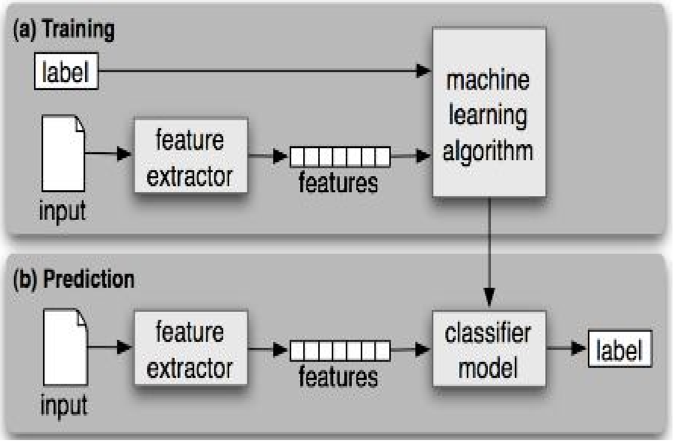

Figure 4 : Show the training and Testing /Prediction process

3.2. Preprocessing

Data pre-processing is a method in which input data is being processed and converted to a machine learning model. It is the first and important step in the creation of a model for machine learning. By doing this we can improve the efficiency and accuracy of the classifier. It contains 3 steps.Data formatting for ML (structured format) formats. Data cleaning to remove missing variables. Data sampling which reduces the running time and also memory. Data cleaning at this stage involves filtering it based on the following variables: Insufficient Data, Non-Representative Data, Substandard Data, Sampling Noise, and Sampling Bias.

3.3. Tools

1) **Lowercasing**: Text has a wide range of capitalizations that represent the start of sentences or the importance of proper names. For simplicity, the usual solution is to simplify it to a smaller case. By converting the text into lower case letters which will be very helpful tool in preprocessing the data[25] for eg: some words like "US" to "us", which the change the whole meaning of the sentence.

2) **Stop Word Removal**: Stop words are a collection of widely used words in English such as "a," "the," "is," "are," and so forth. Such words have significant meaning and in many data analysis activities these must be omitted from documents. For eg IN:['How', 'did', 'he', 'try', 'to', 'escape', 'from', 'the', 'flight', 'accident', 'when', 'suddenly', ',', 'a', 'truck', 'came', '.']

OUT: ['try', 'escape', 'from', '', 'flight', ',', 'truck'', '.'].

3) **Noise Removal:** Noise reduction is the reduction of character digits and code fragments which may conflict with the interpretation of data. Noise elimination can also be achieved by different means, including punctuation reduction, special elimination of character numbers, reduction of html coding, domain-specific[26] removal of keywords, removal of source code, etc.

4) **Tokenization:** It is another very important preprocessing tool in which a sentence can be divided into words. IN: "The mobile what I purchased last month ie apple is very good camera quality is excellent."

OUT:[‘the’,’mobile’, ‘what’,’I’,’ purchased’,’ last’,’ month’,’ie’, ‘apple’.’ Is’, ‘very’,’ good’, ‘camera’,’ quality’,’ is’, ‘excellent.’, '.']

5) **Part of speech**: POS is the combination of rule based or stochastic based method ie it gives the word parts of speech for eg:

IN:[‘The’,’mobile’,’I’,’purchased’,’looks’,’very’,’cool’,’and’,’the’,’camera’,’quality’,’is’,’excellent’]

OUT:[('The', ‘DET'’), (‘Mobile', 'JJ'),(‘I’,’PRP’)(‘Purchased’,’VBD’), ('look', 'VBZ'), (‘Very', 'RB'), (‘Cool', 'JJ'), (‘and’, ‘CC’),(‘Camera', 'NN'), (‘Quality', 'NN'), ('is’, ‘VBZ’), (‘Excellent', 'JJ'),].

6) **Stemming**: Stemming is the process of eliminating suffixes, prefixes, infixes, circumfixes from a word in order to obtain its root word. It converts each and every word into its root word. For eg:”universe” and “university” reduces to its root word ie “univers”.

7) Lemmatization: It is related to stemming which is capable of capturing lemma-based canonical forms. By using lemmatization we can achieve better results. For eg: The lemmatized form of leafs is: “leaf”. The lemmatized form of leaves is: “leaf”

3.4 Overview of CUDA

CUDA is given as Compute Unified Device Architecture. It is an extension of C language, which adds some predefined library functions to access GPU. CUDA is an API architecture developed by Nvidia which works on the principle called parallel computing which increases computing performance. The other principle is a multi-core processor built for vast chunks of data that appear to be unreliable in CPUs. A GPU consists of several cores and each clock speed is slightly slower than one CPU clock. GPUs concentrate on huge parallel programme execution and efficiency. CUDA works on modern nVidia cards (Quadro, GeForce, Tesla). CUDA code must be compiled using nvcc compiler. The compiler generates both instructions for host and GPU (PTX instruction set), as well as instructions to send data back and forwards between them. Fig 8 shows comparison of CPU/GPU

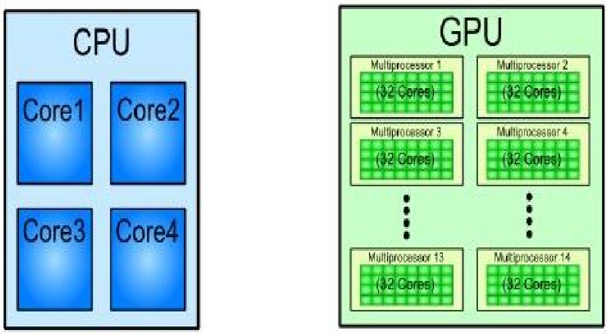

Figure 5: Shows the architecture of CPU VS GPU

In GPU computing we can’t able to access the memory directly and CPU cant access GPU memory then in such situation we have to copy the data explicitly with the help of some predefined libraries ie, 1)

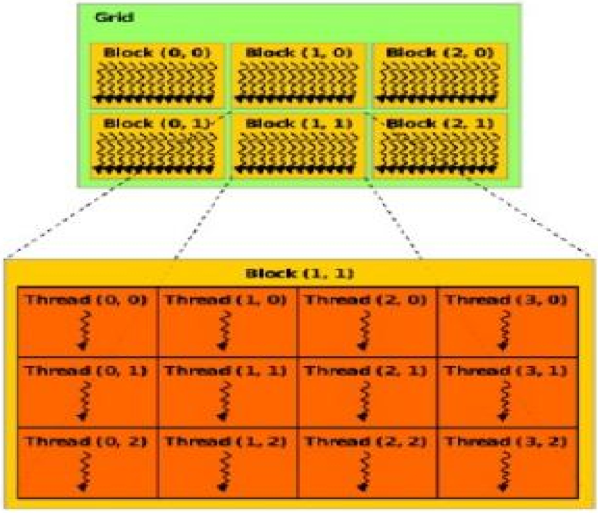

Figure 6: How the Threads are grouped into a Block

4. PROPOSED METHOD

The main purpose of this paper is to predict the accuracy of the classifier using GPU computing on Amazon Unlocked Mobile data i.e., which is based on the feedback given by the product customer in online social network which uses features like effectiveness and specificity to enhance our prediction. In this paper we have taken the reviews or messages about a particular product feature and judge whether it is positive, negative or neutral[27]. To perform this we have applied Vader’s Sentimental Classification for the evaluation of text and phrase levels data. Once the polarity is classified which is represented in the form of binarization i.e. in the form of zero’s and one’s and once the test data was trained then we are generating confusion matrix and calculating Precision,Recall,F1score for finding the accuracy of the data which is done on GPU computing. For this we used 3 classification algorithms [31] (1) CUDABB Algorithm (2) MMDBM (3)SLIQ. Combinations usually are used to test bias and then to determine the method of learning [8].

Bagging: Bagging is a process that incorporates the outcomes of various models (e.g. all decision trees) in order to produce a detailed outcome. For this function, bootstrapping is a sampling strategy in which we establish subsets of analysis from the initial dataset, with a substitution. The size of the subsets is the same as that of the initial set. From the original dataset, multiple subsets are created, selecting substitution observations. For each of these subsets, a base model (weak model) has been created. The models operate in parallel. The final predictions are calculated by a mixture of all the predictions[29].

Boosting: Booting is a sequential operation. In which the errors of the previous model are replaced for the subsequent one. In this paper the hybrid algorithm which is the combination of both Bagging and Boosting is combined i.e. a blended algorithm. Because of its hybrid characteristics, this algorithm produces a classifier that eliminates bias at each training level, while simultaneously regulating the over-fit [17].

CUDA Bag-Boost Algorithm in GPU

```
Input:    A is an array contains n attributes A = {x1, x2,........, xn}  given in parallel.
Output: Decision tree construction and Accuracy of the model.
    1.  Initialize threads in Graphical Processing Unit.
    2.  Generate data values arbitrarily in the dataset.
    3.  Move the data from Graphical Processing Unit device to CPU host (cudaMemcpy) move the values in arrays (binarization).
    4.  Copy these values into Graphical Processing Unit device and apply sorting to arbitrary data from the data sets inside the device GPU.
    5.  Copy these values from CPU host and find the midpoint for each and every attribute.
    6.  Arbitrarily select X features from total Y features where X < Y.
    7.  From the above obtained X features calculate the node Z using best split.
    8.  Next step split these nodes into subnodes ("child nodes") nodes using the best split.
    9.  Repeat the above steps until W number of nodes has been reached.
    10. Build forest by repeating the steps from 1 to 9 for K number of times to create L number of trees.
    11. callXGB()
    12. model=xgb.XGBClassifer(random  state=1,learning rate=0.01)
    13. model._t(x train, y train)
    14. model. score(x test, y test)
    15.  return model score
    16. Transfer the model score from CPU to Graphical Processing Unit device, classify the data, compute the node count and class count.
    17. Send the result obtained from GPU to CPU host, i.e. accuracy of the classifier.
    18. End process.
```

5. IMPLEMENTATION AND RESULTS

The classification algorithms used like SLIQ, MMDBM and CUDABB. Further when executed in CUDA programming all the algorithms are compared in terms of processing time. In which GPU mining takes less processing time when compared to CPU .So The GPU mining algorithm which when tested on various threads shows that GPU computing is much faster than CPU computing[30]. To check the effectiveness of the algorithm, we are going to apply the algorithm to a randomly generated Amazon Unlocked Mobile dataset, where the task is to predict the star rating of a mobile which is based on the classification done on random database [3].

The implementation of the algorithm can be divided into seven phases and is described as follows:

In CUDA connection between the database and CUDA is not possible.So, the data for classification has to be generated randomly.

The first phase involves generating random data, which will take care by CUDA using a built-in function called curand. In generating a random data CUDA programming is very fast, i.e, the time taken for generating one lakh of data has took 0.05 second.In the second phase we have to fix the number of attributes, type of the attributes and label of the attributes.

The third phase of the implementation the attributes are represented in the form of a decision tree [21,22]. Here decision tree is a binary tree, because we classify only completed data sets.In the fourth phase sorting the generated values and determining the split point for each numerical attribute is done.The fifth phase of implementation includes in modeling the decision tree-based classification rules which are used in GPU programming. The sixth phase of the implementation involves in memory allocation for the data in the device (GPU). This is performed by a built-in function called cudaMalloc. After that copy the obtained data to the GPU device. Which is performed by a built-in function called cudamemcpy. This method has to be used twice i.e, copying the data from host(CPU) to the device(GPU) and then copying thSe results from device(GPU) to host(CPU).The final phase of the implementation involves in finding the accuracy of the results in GPU.The classification takes place when only 128 threads are launched for one lakh of results. This is because the data is autonomous. The number of data supplied refers to the number of threads initiated and the process is finished within microseconds. Table 1 shows the different numbers of threads and times.

Table 1: Time taken by varying the number of CUDA Threads.

Number of threads	Time Taken
128	5.45
256	4.34
512	3.25
1024	2.82

Table 2: Acceleration Ratio time for Classifying Records using CUDABB Algorithm

CUDABB GPUs Times	No.of. Records Sec / 10000	No.of. Records Sec / 30000	No.of. Records Sec / 50000	No.of .Records Sec/ 70000	No.of. Records Sec / 100000
Classification Time	0.535	1.015	1.620	2.045	2.684
CPU Time	0.710	1.130	1.740	2.3500	2.900
GPU Time	0.550	1.010	1.640	2.230	2.490
Acceleration Ratio	1.296	1.118	1.064	1.054	1.16491

5.1. Acceleration Ratio for GPU

To calculate GPU Performance: To find the acceleration ratio of Graphical Processing Unit.

(GPU) ie (speed-up time) which is represented with a

$$\gamma = \frac{t_{CPU}}{t_{GPU}}$$

symbole γ is defined as where tCPU the total processing time on the CPU, tGPU is the total processing on the GPU.

CPU Time = Generate the random Values + Sorting Time + Classification Time.

GPU Time = Data transfer from Host to Device and Device to Host.

Acceleration Ratio = CPU computation Time / GPU computation Time.

Dataset Description: Dataset contains around 5, 00,000+ reviews of Mobile phones, collected from Kaggle repository. This is existed in the form of .csv file.

Table 3: Dataset Description

Features	Values
Name of the data set	Amazon.csv
Dataset Url	F:\python\proj\Amazon1.csv
Number of Reviews	413840
Classes	Positive, Negative and Neutral
File Types	CSV

For Evaluations we perform the sentiment analysis using CUDABB approach. The results which are achieved is compared with other 2 classification algorithms like SLIQ,MMDBM. The CUDABag-Boost classifier has achieved high accuracy when compared to the other two. From this it can be concluded that CUDABB algorithm gives more accuracy when compared to the other 2 Machine Learning algorithms which when runned on GPU.

5.2. Measures for performance evaluation:

Precision: It is data retrieval; precision is a metric that quantifies the number of correct positive predictions made [25]. Mathematically:

$$\mathrm{Pr}ecision = \frac{TP}{(TP + FP)}.$$

Recall: This is data retrieval, recall is the fraction between correctly classified positives by the classifier and manual classified positive (true positive + false negative). Mathematically:

$$\mathrm{Re}call = \frac{TP}{(TP + FN)}.$$

F- measure: This is the harmonic mean between precision and recall. Mathematically:

$$F - measure = 2 * \frac{\mathrm{Pr}ecision * \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}.$$

Accuracy: This is calculated as the proportion of true positives, true negatives, false positives and false negatives. The true results from all the given data. Mathematically

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$

G-mean: This is geometric mean of precision and recall and is only the height when both of these measures are high[26]. Mathematically

$$G - mean = \sqrt{\mathrm{Pr}ecision * \mathrm{Re}call}$$

BCR and BER: Balanced Classification Rate and Balanced Error Rate (BER) is correctly balanced rate on each class and balanced error rate is the average of the errors on each class.

$$BCR = \frac{1}{2}\ (Sensitivity + Specificity).$$

$$BER = 1 - BCR.$$

The result of the proposed methods is evaluated and compared by using the confusion matrix, sensitivity, specificity, precision, recall, F-measure, G-mean, BCR and accuracy.

Table 4: Shows the accuracy of MMDBM Algorithm**.**

Star rating	Precision	recall	f1-score	support
1	0.84	0.81	0.80	8138
5	0.90	0.94	0.85	23577
Microavg	0.86	0.87	0.90	31715
Macroavg	0.88	0.89	0.88	31715
Weightedavg	0.90	0.90	0.90	31715

Table 5: Shows the accuracy of SLIQ Algorithm**.**

Star rating	Precision	recall	f1-score	support
1	0.83	0.87	0.85	8138
5	0.92	0.96	0.94	23577
Microavg	0.83	0.83	0.83	31715
Macroavg	0.78	0.69	0.71	31715
Weightedavg	0.88	0.88	0.88	31715

Table 6: Shows the accuracy of CUDA Bag-Boost Algorithm.

Star rating	Precision	recall	f1-score	support
1	0.86	0.78	0.80	8138
5	0.94	0.92	0.94	23577
Microavg	0.94	0.90	0.94	31715
Macroavg	0.88	0.84	0.88	31715
Weightedavg	0.94	0.94	0.94	31715

Table 7: Comparison of CUDABB with supervised learning methods for Amazon Mobile Phone dataset.

Methods	Precision	Recall	F-measure	G-means	BCR	BER	Accuracy
Proposed CUDABB	0.945	0.944	0.944	0.931	0.944	0.046	0.944
MMDBM	0.891	0.909	0.898	0.858	0.901	0.099	0.900
SLIQ	0.886	0.877	0.879	0.825	0.880	0.120	0.877

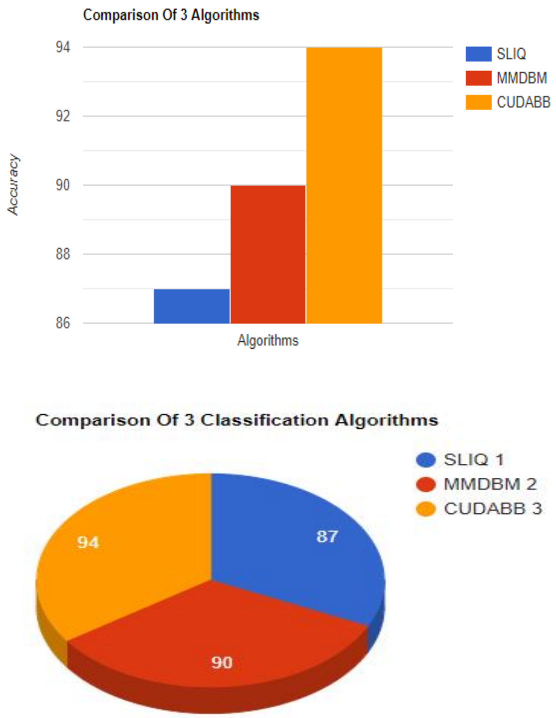

Figure 7: Accuracy Measure comparison result of 3 Machine Learning Algorithms.

6. CONCLUSION AND FUTURE WORK

The proposed CUDABB algorithm which when implemented with large number of attributes in GPU parallel computing for classifying dataset. Sentiment analysis shows that positive or negative or neutral sentiments is established among the reviews and in terms of emotion The above said algorithms can be implemented for aspect classification and polarity identification of product using machine learning algorithms The obtained results are compared with SLIQ and MMDBM using GPU which results in acceleration ratio time. GPU Mining increases the performance with less processing time. The proposed method achieves better accuracy of 94% with 94.50% precision, 93.95% recall, 94.21% F-measure, 94.13% BCR and 9.67%BER. In future research, the proposed algorithm can be applied in various real time applications such as Banking, Bio-medicine, and Big Data Analytics[28].

7 ACKNOWLEDGEMENT

“N.-Anabzhagan's research has been financially supported jointly by RUSA-Phase-2.0 grant sanctioned vide-letter-No.F-24-51/2014-U,Policy(TNMulti-Gen),-Dept.-of-Edn.-Govt.-of-India,-Dt.- 09.10.2018,-UGC-SAP-(DRS-I) vide letterNo.F.510/8/DRS-I/2016(SAP-I) Dt.23.08.2016,DST PURSE-2nd-Phase-programme-vide-letter-No.-SR/PURSE-Phase-2/38-(G)-Dt.-21.02.2017-and-DST- (FIST level-I)657876570 vide letter-No.SR/FIST/MSI/2018/17-Dt.-20.12.2018.”

REFERENCES

[1] Siva Kumar Pathuri, N.Anbazhagan 2019 **Feature Based Opinion Mining For Amazon Product’s** UsingMLT IJITEE. 8(11)

[2] SivaKumarPathuri and N. Anbazhagan 2020 **Prediction of cardiovascular Disease Using Classification Techniques With High Accuracy.** JARDCS.12(02).

[3] A.Ejaz, Z.Turabee, M.Rahimand S.Khoja 2017 **Opinion mining approaches on Amazon product reviews:A comparative study.** InternationalConferenceonInformationandCommunicationTechnologies(ICICT),Karachi,pp173-179.

[4] Sivakumar S, Periyanagounder, G., Sundar S, A **MMDBM classifier with CPU and CUDA GPU computing in various sorting procedures.** International Arab Journal of Information Technology. 14(6). pp 897-906.

[5] S Anjali Deviand S Siva **Kumar2020A hybrid document features extraction with clustering based classification framework on large document sets**. International Journal of Advanced Computer Science and Applications.11(7) pp 364-374.

[6] Anila, M., Pradeepini, G.**Least square regression for prediction problems in machine learning using R** ,2018”International Journal of Engineering and Technology(UAE),PP:960-962.

[7] MsKrantiGhagand, Dr.KetanShah 2013 **Comparative Analysis of the Techniques for Sentiment Analysis** ICATE 2013 Paper IdentificationNumber124.

[8] A.J.Singh 2017 **Sentiment Analysis :A Comparative Study of Supervised Machine Learning Algorithms Using Rapidminer** IJRASET. 5(Xi). pp 80–89.

[9] Baig, M.M, Sivakumar, S, S R Nayak 2020 **Optimizing Performance of Text Searching Using CPU and GPUs Advances in Intelligent Systems and Computing**, 1119. pp 141-150.

[10] Sivakumar, S, Nayak, S.R, Vidyanandini, S, Kumar, J.A, Palai, G 2018 **An empirical study of supervised learning methods for breast cancer diseases,** Optik-International Journal for Light and Electron Optics.175 pp 105–114.

[11] Sreedevi E, Premalatha, V, Sivakumar, S, Nayak, S.R 2019 **A comparative study on new classification algorithm using NASA MDP datasets for software defect detection**. Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019. pp 312-317.

[12] V.PremaLatha, E Sreedevi, Sivakumar, S 2019 **Contemplate on internet of things transforming as medical devices - The internet of medical things**. (IOMT)Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019. pp 276-281.

[13] Shaik Razia, P.SwathiPryathyusha, N.Vamsi Krishna 2018 **A Comparative study of machine learning algorithms on thyroid disease prediction.** International Journal of Engineering and Technology(UAE). 7(2.8). pp 315-319.

[14] ShailRazia, M.R.Narasingarao 2017 **Development and Analysis of Support Vector Machine Techniques for Early Prediction of Breast Cancer and Thyroid**. JARDCS (Journal of Advanced Research in Dynamical and Control Systems). 9(6). pp 869-878.

[15] ShailRazia, M.R.Narasingarao **2017A Neuro computing frame work for thyroid disease diagnosis using machine learning techniques.** JATIT (Journal of Theoretical and Applied Information Technology). 95(9). pp 1996-2005.

[16] Videla, Lakshmi Sarvani, et al. **"Deformable facial fitting using active appearance model for emotion recognition."** Smart Intelligent Computing and Applications. Springer, Singapore, 2019. 135-144.

[17] Videla, Lakshmi Sarvani, et al. **"Modified Feature Extraction Using Viola Jones Algorithm".** Journal of Advanced Research in Dynamical and Control Systems.Volume 10, Issue 3 Special Issue, 2018, Pages 528-538.

[18] Videla, Lakshmi Sarvani and M. Ashok Kumar P. **"Fatigue Monitoring for Drivers in Advanced Driver-Assistance System."** IGI Global, 2020, pp. 170-187.

[19] Atmakur, V.K., Siva Kumar, P.”**A prototype analysis of machine learning methodologies for sentiment analysis of social networks”**, International Journal of Engineering and Technology(UAE), 2018,pp 963-967.

[20] Prabha Selvaraj ,Vijay Kumar Burugari , D. Sumathi ,Rudra Kalyan Nayak ,Ramamani Tripathy**,”Ontology based Recommendation System for Domain Specific Seekers”**,Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) 2019,pp341.

[21] T. Shiva, T. Kavya, N. Abhinash Reddy, Shahana Bano**.”Calculating The Impact Of Event Using Emotion Detection”,** International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7 May, 2019.

[22] Prasanth Kumar1* Gera Pradeepini1. Pille Kamakshi**,” Feature Selection Effects on Gradient Descent Logistic Regression for Medical. Data”,** October 2019,International Journal of Intelligent Engineering and Systems 12(5) pp:278-286.

[23] Chintala, R. R., Narasinga Rao, M. R., \& Venkateswarlu, S. (2018). **A proposal for observing conceived ladies having high risk of premature delivery using WHSN.** International Journal of Engineering and Technology(UAE), 7(2), 53-56. doi:10.14419/ijet.v7i2.32.13524.

[24] Sajana, T., \& Narasingarao, M. R. (2018). **An ensemble framework for classification of malaria disease.** ARPN Journal of Engineering and Applied Sciences, 13(9), 3299-3307.

[25] Chintala, R. R., Narasinga Rao, M. R., \& Venkateswarlu, S. (2018). **Review on the security issues in human sensor networks for healthcare applications.** International Journal of Engineering and Technology(UAE), 7(2.32 Special Issue 32), 269-274.

[26] Deshpande, L., \& Rao, M. N. (2018). **Concept drift identification using classifier ensemble approach**. International Journal of Electrical and Computer Engineering, 8(1), 19-25. doi:10.11591/ijece.v8i1.pp19-25.

[27] Arba Asha Altaye, Dr. J. Sebastian Nixon,**A Comparative Study on Big Data Applications in Higher Education.**Volume 7, No. 12 December 2019,International Journal of Emerging Trends in Engineering Research.

[28] Gunawan Wang, Natanael Alamas, and Marcelina Anggraeni, **The Use of Internet of Things and Big Data to Improve Customer Data in Insurance Company.** Volume 7, No. 12 December 2019 International Journal of Emerging Trends in Engineering Research.

[29] Mane, S. U., & NarsingaRao, M. R. (2019). **Large-scale compute-intensive constrained optimization problems: GPGPU-based approach.** doi:10.1007/978-981-13-0589-4_54.

[30] Swetha, K., & Narasinga Rao, M. R. (2016). **Dynamic searchable encryption over distributed cloud storage**. Asian Journal of Information Technology, 15(23), 4763-4769. doi:10.3923/ajit.2016.4763.4769.