Sentiment Analysis in Twitter Dataset using Classification Algorithm

*Elavarasi D, #Dr. Kavitha R

Department of Computer Science and Engineering

*Assistant Professor, Mount Zion College of Engineering and Technology, Pudukottai, Tamilnadu, India.

#Professor, Velammal College of Engineering and Technology, Maduari, Tamilnadu, India.

*elavarasijournal@gmail.com

Abstract - Sentiment analysis is an analyzing data mining of social media. Opinion mining is the process of tracing opinions, views or suggestions of a particular twitter dataset. Today it's all about giving opinions as positive, negative and neutral. From this able to get knowledge from sentiment analysis of social media. By means of these online applications huge number of opinions is given by the user. The reviews of twitter dataset which gives the success level of the twitter. There are many algorithms have been used to find the opinion in sentiment analysis. There are some aspects of textual content, which form equally valid selection criteria. This paper presents the sentence-level opinion mining classification using CART and C4.5 algorithm.

Keywords: Sentiment analysis, Sentence level, Opinion mining, CART and C4.5 algorithm.

I. INTRODUCTION

1.1 Data mining- Data mining is a mining knowledge from large amount of data. Data mining uses sophisticated mathematical algorithms to fragment the data and evaluate the probability of future events. Data mining is also called as Knowledge Discovery in Data (KDD).

1.2 Sentiment analysis- Sentiment analysis is learning of people's emotions, views, attitude, and opinions. It is also called an opinion mining. Sentiment analysis identifies the sentiment articulated in a text then analyzes it. So sentiment analysis to find the opinions, show the sentiment and classify the polarity. Opinion mining used to study the sentiments spoken by people on the internet through reviews. Opinion mining is a type of Natural Language Processing (NLP) for tracking the mood of the public through a particular product. There is a huge literature on sentiment analysis (Pang and Lee, 2008; Liu, 2012), with particular interest in determining the overall sentiment polarity of a document. For example, movie reviews help new users to decide whether the movie is watch or not. Though, the huge numbers of review become information overload absence of automated methods for computing their sentiment polarities.

There are three levels in sentiment analysis. Document-level, Sentence-level, and aspect-level.

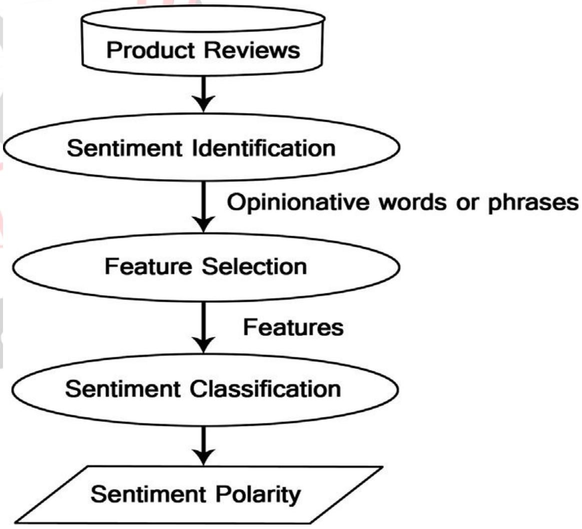


Figure1.1: Sentiment Analysis Process

1. Document-level: It classify the document as positive, negative or neutral. It is known as document-level sentiment classification.

2. Sentence-level: It classify the sentences as positive, negative or neutral. It is known as sentence-level sentiment classification.

3. Aspect-level: It classifies the sentiment to the specific aspects of entities. It is known as aspect-level sentiment classification.

1.3Document-level Sentiment Analysis- Document-level sentiment analysis aims to organize the view text as

DOI : 10.35291/2454-9150.2019.0450

expressing an optimistic or pessimistic opinion or sentiment. In Document-level sentiment classification to classify a textual analysis which is specified on a particular topic. The task is also commonly known as the document-level sentiment classification for the reason that it considers the document as the basic information unit.

II. LITERATURE SURVEY

The most prominent work done by Walaa Medhat et al [1] the authors have discussed the feature selection techniques and sentiment classification techniques during information beside through their connected articles referring to several originating references. These fields include Emotion Detection, Building Resources and Transfer Learning. The accuracy percentage of context based SA, which is called as domain dependent data is more than that of domain independent data.

Research in the opinion mining of movie reviews at document level proposed by Richa Sharma et al [2] has mentioned the planned work is directly connected to the Minqing Hu and Bing Liu work on mining and shortening purchaser reviews. The proposed system divided into phases such as (i) Data Collection (ii) POS Tagging (iii) Extracting opinion words and seed list preparation and polarity detection and classification. The conduct experiment outcome shows that the Document based Emotion Orientation System at hand well between outlays to the picture meadow as compared to 'AIRC Sentiment Analyzer'. Upcoming structure makes the truth of 63%.

In the paper Document-level sentiment classification: An empirical comparison between SVM and ANN proposed by Rodrigo Moraes et al[3] focused on comparing SVM and ANN in terms of requirements to achieve better classification in accuracies. Their experiments evaluated both methods as a function of selected terms in a bag of words approach. Although the accuracy between them has never exceeded by 3%, ANN have achieved the greatest organization precision in every datasets. However their results indicated that SVM technique is less affected by noisy terms than ANN, when the data imbalance increases.

Researches in Better document level sentiment study from RST Discourse Parsing planned by Parminder Bhatia et al[4] has presented two different ways of combining RST discourse parse with sentiment analysis. The methods are simple and can use in combination with an "off the shelf" dissertation parses. They consider the following two architectures (i) Reweighting the contribution of each discourse unit and (ii) Recursively propagating sentiment up through the RST parse. Both the construction can be used in grouping with also a lexicon-based sentiment analyzer or a educated classifier. They evaluated on the Pang and Le data and consider only lexicon-based sentiment analysis, obtaining document level inaccuracies between 65% (for baseline) and 72% (for their best discourse-augmented system).

Identifying high impact substructures for convolution kernel in document level sentiment classification proposed by ZhaopengTu et al[5] evaluated diverse linguistic structures determined as difficulty kernels for the document level sentiment classification trouble, to use syntactic structures without defining explicit linguistic rules. They explored Subset Tree (SST) and Partial Tree (PT) kernels for component and reliance parse tress correspondingly. The best performance had achieved by combining VK and DW kernels, gaining a significant improvement of 1.45 point in accuracy.

The paper, "Retrieving topical sentiment from online document collections" planned by Matthew Hurst et al [6] has offered a lightweight but robust move toward to combining topic and polarity. The method they used for analyzing a paper was individual language by responsibility a fine-grained NLP based textual study and machine learning classification based approach. This paper strikes away a central point view connecting these two approaches and argues for a union of polarity and topically. The evaluated three aspects of their move toward (i)the presentation of the subject classifier on sentences (ii)the routine of the polarity detection system and (iii)the hypothesis that polar sentences are on area include polar language.

Wei-Hao Lin et al[7] a new problem of learning to name the outlook from which a text written at the document and sentence levels. A large amount of the document's perspective articulated in word procedure, and arithmetical knowledge algorithms such as SVM and Naive Bayes duplicate are clever to profitably determine missing the word patterns that articulate author point of view with high accuracy.

AinurYessenalina et al[8] covert variable structured model used for the document sentiment classification assignment. These models do not rely on sentence-level interpretation, and educated jointly to directly optimize document-level precision.

In the paper, "computing sentiment polarity of texts at document and aspect levels proposed by Vivek Kumar Singh et al[9] proposed two methods such as lexicon based and heuristic based scheme. The authors evaluate presentation of four different sentiment analysis schemes on six diverse datasets absent of the four implementations, two be machine learning classifies since the left over two are lexicon-based methods.

In the paper[10] objective of this paper shows that it is to determine the polarity of the movie reviews or criticisms at the document level. The results produced by the system are shortened and supportive for the client in decision making. Experimental outcomes state that the Document-

DOI : 10.35291/2454-9150.2019.0450

based Sentiment placement system implement healthy in this domain. Opinion mining is very substantial these days from the general man to a business man , everyone is needy on the web. The opinions communicated on the web benefits the users to limit which creation or movie is good for them and it helps the businessman to regulate what the clients thinks about their products . So, it is compulsory to mine this large amount of criticisms and organize them, so it is helpful for them to read and yield conclusions.

This paper [11] covers our participation in the ABSA (Abstract based –sentiment Analysis task of semi level the ABSA task involves of 4 subtasks. For every subtask we suggest both embarrassed and unhindered attitude. The controlled descriptions of our classification are established innocently on machine learning procedures. The proposed approaches accomplish very good results. The constrained varieties were always above average, habitually by the large boundary the unconstrained versions were ranked midst the greatest systems.

The objective of this paper [12] to determine the polarity of the customer reviews of mobile phones at aspect level. The system performs the aspect-based opinion mining on the given reviews and the feature wise brief results created by the system will be supportive for the user in enchanting the decision. Cautious technique results or outcomes indicate that the Aspect based Sentiment orientation system that is capable well and has completed the accurateness of 67%.

In this paper [13] design a deep learning model to analyze the aspect based sentiments and demonstrates competitive or better performance comparing to the results of SemEval'15 in all subtasks. Propose a novel approach to connecting sentiments with the corresponding aspects based on the constituency parse tree. This model also shows promising performance in an unseen domain. In the future work, we are interested in testing the model on other datasets and evaluating the performance of transfer learning. It is also like to explore many models in aspect of prediction by using adaptive thresholds.

In the [15] a novel method to deal with the problems. An augmented lexicon- based method specific to the twitter data was first functional to accomplish sentiment analysis. Complete Chi-square test on its output, further blinkered tweets could be acknowledged. A binary sentiment classifier is then skilled to disperse sentiment polarizations to the freshly- identified blinkered tweets, whose preparation data is delivered by the lexicon-based method. Realistic experiments display the proposed method is high – effective and hopeful.

III. PROPOSED SYSTEM

A sentence based opinion mining classify the document as positive, negative and neutral. It is also handled using CART, C4.5, Navie bayes require a high amount organize review as expressing optimistic or pessimistic opinions at sentence level. time-based data analysis. The R language is widely used among data miners for developing statistical

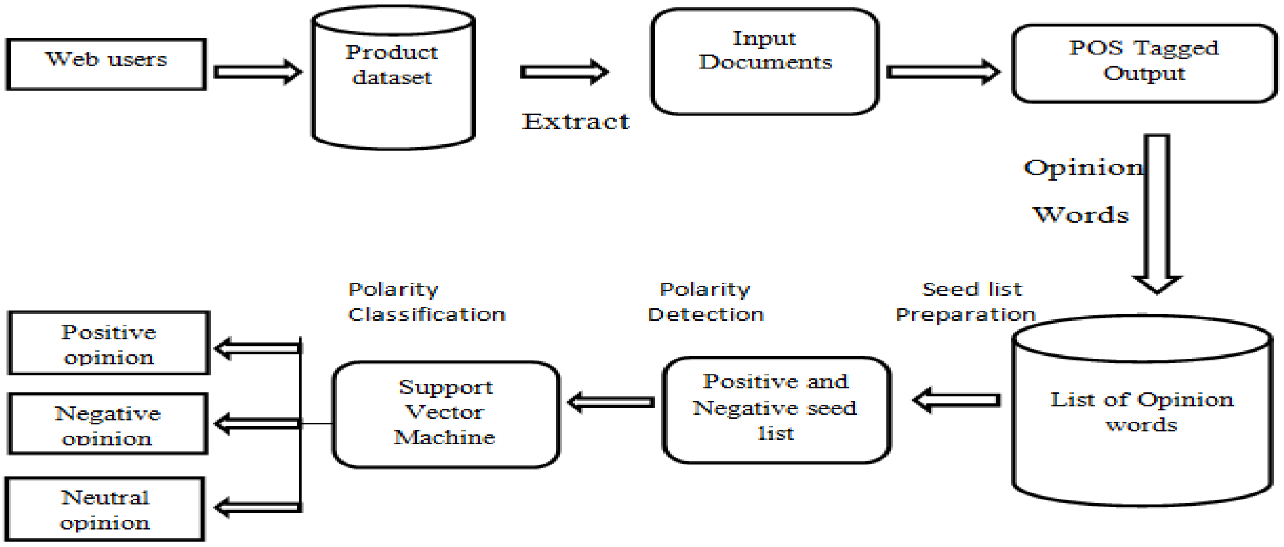


Figure 3.1 Proposed Works

pessimistic opinions at sentence level. Classification model gives the best accuracy. The main goal is to retrieving documents by subject and other content access system. The two standard sentiment analysis datasets shows improvement in performance. The classification task is well modeled by jointly solving an extraction subtask. The experiment uses Sentiment analysis twitter dataset obtained from UCI machine learning repository. The data set consists of total 1038 instance 2 attributes; In this experiment, 2 attributes are used.

DOI : 10.35291/2454-9150.2019.0450

3.1 R-PROGRAMMING TOOL

This is written in C and FORTRAN, and allows the data miners to write scripts just like a programming language/platform. Hence, it is used to make statistical and analytical software for data mining. It supports graphical analysis, both linear and nonlinear modeling, classification, clustering and software and data analysis. non linear modeling, classification, clustering and time-based data analysis. The R language is widely used among data miners for developing statistical software and data analysis.

3.2 POS TAGGING

Once gathering the reviews, they are shown to the POS tagging element where POS taggers that tag all the words of the sentences to their suitable part of speech tag. POS tagging is an essential segment of opinion mining, it is compulsory to fix the structures and opinion words from the reviews.

POS tagger is used to tag all the words of reviews or criticisms.

3.3 EXTRACTING OPINION WORDS

Basically few of the general opinion words along with their polarity that is stored in the seed list. All the opinion words are mined from the tagged (pos) output, The extracted opinion words that are matched with the words stored in the seed list. If the word is not found in the seed list then the synonyms are determined with the help of word net. Every synonym is harmonized with the words in the seed list if any accorded synonym then the mined opinion word is stored with the similar polarity in the seed list.

IV. METHODOLOGY

4.1 CART

Globally-optimal classification tree analysis [16] (GO-CTA) (also called hierarchical optimal discriminant analysis) is a sweeping statement of optimal discriminant analysis that it is used to recognize the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a given dataset which consisting of categorical and continuous variables. The output of Hierarchical Optimal Discriminant Analysis (HODA) it is a non-orthogonal tree which combines categorical variables and also the cut points for continuous variables that yield utmost predictive accuracy, an evaluation of the exact Type I error rate and an assessment of probable cross-generalizability of the statistical model. The HODA may be thought of as an overview of Fisher's linear discriminant analysis. An alternative to ANOVA (analysis of variance) and regression analysis is an optimal discriminant analysis, which also attempts to express one dependent variable as a linear combination of other features or measurements. However, regression analysis ANOVA and gives a dependent variable that is a numerical variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable.

Classification and regression trees (CART) [16] are a non-parametric decision tree learning technique that gives either regression trees or classification, depending on whether the dependent variable is categorical or numeric, correspondingly. Decision trees are formed by a group of rules based on variables within the modeling collected data set: The first step is rules-based on variables' values are designated to get the finest fragment to differentiate observations based on the dependent variable.

The second step after the rule is designated and fragments a node into two parts, the same procedure is applied to each "child" node (i.e. it is a recursive procedure)

The third step is fragmenting stops when CART detects no further gain can be made, or some pre-set stopping rules are met.

The fourth step in each branch of the tree ends in a terminal node. In each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

Gini impurity [16] used by the CART algorithm for classification trees, it is used to measure a randomly chosen component from the dataset would be erroneously labeled if it was randomly labeled according to the distribution of labels in the subset. By summing the probability p_i of an item with label i being chosen times the probability

$$\sum_{k\neq i} p_k = 1 - p_i$$

of a mistake in categorizing that item by this way, Gini impurity can be computed. It reaches it's minimum (zero) value when all cases in the node fall into a single target category.

In order to compute Gini impurity for a given dataset of items with J classes, suppose $i \in \{1, 2, \ldots, J\}$, and let pi be the fraction of items labeled with class i in the set.[16]

$$\mathrm{I}_G(p) = \sum_{i=1}^{J} p_i \sum_{k\neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J}(p_i - {p_i}^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} {p_i}^2 = 1 - \sum_{i=1}^{J} {p_i}^2$$

4.2 C4.5 ALGORITHM

C4.5 is an algorithm used to make a decision tree developed by Ross Quinlan.C4.5 is commonly noted as a applied mathematics classifier. C4.5 constructs a decision tree from a set of training data in the same way as ID3, using the idea of information entropy. The f is a set which is a classified sample from the training data. In each sample Si consists of a p-dimensional vector where the xj represent attribute values of the sample, as well as the class in which si falls. At each node of the tree, C4.5 chooses the attribute of the data that most efficiently splits its set of samples into subsets enriched in one class or the other. The cacophonous criterion is that the normalized info gain (difference in entropy). The attribute with the best normalized info gain is chosen to create the choice. This algorithm has a few base cases. All the samples in the list belong to the same class. When this happens, it merely creates a leaf node for the choice tree speech to decide on that class.

None of the features provide any information gain. In this case, C4.5 creates a call node in a higher place the tree victimisation the first moment of the class. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

V. RESULTS AND DISCUSSION

Twitter Dataset

The experiment uses Sentiment analysis twitter dataset obtained from UCI machine learning repository. The data set consists of total 1038instance 2 attributes; in this experiment 2 attributes are used.

Number of Instances: 1038

Number of Attributes: 2 (all nominally valued) . Here sample dataset is given

The attributes are as follows:

User Name	Keyword Repetitions	Follower/Following ratio	Sentiment	Comments
Joel Comm	1	1.009835244655297	1	Now all @Apple has to do is get swype on the iphone and it will be crack Iphone that is
Vincent Boucher □	3	1.393611474853291	1	@Apple will be adding more carrier support to the iPhone 4S (just announced)
William Tincup	3	0.9262506281547773	1	Hilarious @youtube video - guy does a duet with @apple s Siri Pretty much sums up the love affair
Aaron Carter	1	1.712376685139331	0	@RIM you made it too easy for me to switch to @Apple iPhone. See ya!
BILLIONAIRE PR GIRL□	1	1.414292597922085	1	I just realized that the reason I got into twitter was ios5 thanks @apple
Kim Garst	1	1.633711779396025	1	Im a current @Blackberry user little bit disappointed with it! Should I move to @Android or @Apple @iphone
#DeepLearning #App □	1	1.809501725816444	1	The 16 strangest things Siri has said so far. I am SOOO glad that @Apple gave Siri a sense of humor!
Andrea Feczko	1	1.071963734489032	1	Great up close & personal event @Apple tonight in Regent St store!
Stephen Stephan	4	1.00122159366598	1	From which companies do you experience the best customer service aside from @zappos and @apple?
Neechi	1	2.38539299201351	0	Just apply for a job at @Apple hope they call me lol
#Talk2Me	4	2.724642740108567	1	RT @JamaicanIdler: Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested where to find whores and where to h ...
IG @MDoTMancini	3	0.9730110392082223	1	Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested where to find whores and where to hide a body lolol
Kirby Ellis	1	1.306397437292985	1	RT @PhillipRowntree: Just registered as an @apple developer... Heres hoping I can actually do it... Any help greatly appreciated!
Lori Ruff	1	1.0004888939589	1	Wow. Great deals on refurbed #iPad (first gen) models. RT: Apple offers great deals on refurbished 1st-gen iPads @Apple
Ron Edmondson	1	1.157284355751802	1	Just registered as an @apple developer... Heres hoping I can actually do it... Any help greatly appreciated!
Marcus2braids	1	1.006593172875829	0	Just registered as an @apple developer... Heres hoping I can actually do it... Any help greatly appreciated!
AboveAverageClothing	1	0.9436403977014268	1	! Currently learning Mandarin for my upcoming trip to Hong Kong. I gotta hand it to @Apple iPhones & their uber useful flashcard apps
im n DALLAS	2	1.011449842152382	1	Come to the dark side @gretcheneclark: Hey @apple if you send me a free iPhone I will publicly and ceremoniously burn my #BlackBerry.
Samantha	3	1.16577645311378	1	Hey @apple if you send me a free iPhone (any version will do) I will publicly and ceremoniously burn my #BlackBerry.
The Product Poet	1	1.513204216523073	1	Thank you @apple for Find My Mac - just located and wiped my stolen Air. #smallvictory #thievingbastards
STEPH (OnHol) // ifb	3	1.318784508120743	1	Thanks to @Apple Covent Garden #GeniusBar for replacing my MacBook keyboard/cracked wristpad during my lunch break today out of warranty.

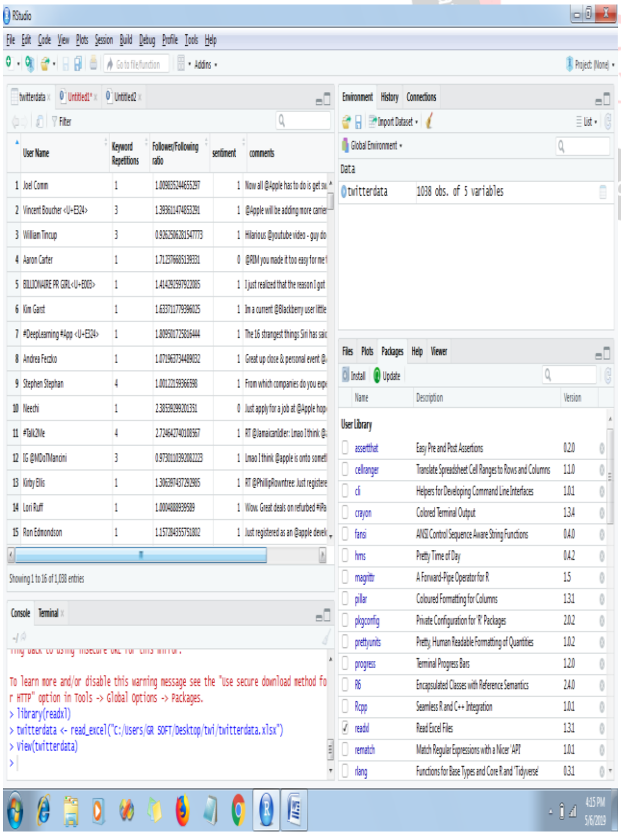
DOI : 10.35291/2454-9150.2019.0450

Calvin Lee	1	0.96983730300040291	1	@DailyDealChat @apple Thanks!!
Elnor Bracho	1	2.119291371755982	1	iPads Replace Bound Playbooks on Some N.F.L. Teams@apple @nytimes
J'Corey Lamar	1	1.248638202533527	0	@apple..good ipad
Maliachi Broadwater	10	1.000248726627808	1	@apple @siri is effing amazing!!
IG: @JabariStaffz	3	1.595399843651938	1	Amazing new @Apple iOs 5 feature. jatFVfpM
Gordon Tredgold	1	2.008082840638036	1	RT @TripLingo: Were one of a few Featured Education Apps on the @Apple **Website** today sweet!
Bill Hibbler	1	1.180089809516912	1	Were one of a few Featured Education Apps on the @Apple **Website** today sweet!

Table 4.1

VI. RESULT ANALYSIS

A sentence based opinion mining classify the document as positive, negative and neutral. It is also handled CART and Trees require a high amount organize review as expressing optimistic or pessimistic opinions at sentence level. Classification model gives the best accuracy but it requires more training time than K-Means the main goal is to retrieving documents by subject and other content access system. The two standard sentiment analysis datasets shows improvement in performance. The classification task is well modeled by jointly solving extraction subtask. All the documents are manually calculated and it is compared with the proposed system to check the performance efficiency of the system. The opinion is applied to the system by using techniques such as CART and C 4.5 (decision tree algorithm). Finally the results of techniques of the proposed system have shown that the Mean, Median, Execution in seconds.



	User Name	Keyword Repetitions	Follower/Following ratio	sentiment	comments
1	Joel Comm	1	1.0098352446655297	1	Now all @Apple has to do is get sw
2	Vincent Boucher <U+E324>	3	1.393611474853291	1	@Apple will be adding more carrier
3	William Tincup	3	0.9262506281547773	1	Hilarious @youtube video - guy do
4	Aaron Carter	1	1.712376685139331	0	@RIM you made it too easy for me t
5	BILLIONAIRE PR GIRL<U+E003>	1	1.414292597922085	1	I just realized that the reason I got
6	Kim Garst	1	1.633711779396025	1	In a current @Blackberry user little
7	#DeepLearning #App <U+E324>	1	1.809501725816444	1	The 16 strangest things Siri has said
8	Andrea Feczko	1	1.071963734489032	1	Great up close & personal event @
9	Stephen Stephan	4	1.00122159366598	1	From which companies do you expe
10	Neechi	1	2.38539299201351	0	Just apply for a job at @Apple hop
11	#Talk2Me	4	2.724642740108567	1	RT @JamaicanIdler: Lmao I think @
12	IG @MDoTMancini	3	0.9730110392082223	1	Lmao I think @apple is onto somet
13	Kirby Ellis	1	1.306397437292985	1	RT @PhillipRowntree: Just registere
14	Lori Ruff	1	1.0004888939589	1	Wow. Great deals on refurbed #iPa
15	Ron Edmondson	1	1.157284355751802	1	Just registered as an @apple devel

Showing 1 to 16 of 1,038 entries

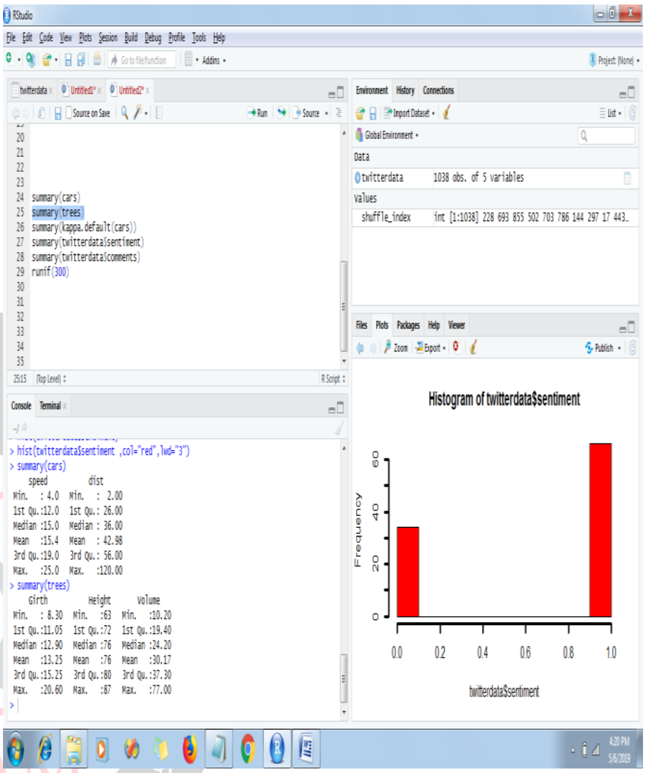
```
Data
twitterdata    1038 obs. of 5 variables
```

	Name	Description	Version
User Library			
□	assertthat	Easy Pre and Post Assertions	0.2.0
□	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
□	cli	Helpers for Developing Command Line Interfaces	1.0.1
□	crayon	Colored Terminal Output	1.3.4
□	fansi	ANSI Control Sequence Aware String Functions	0.4.0
□	hms	Pretty Time of Day	0.4.2
□	magrittr	A Forward-Pipe Operator for R	1.5
□	pillar	Coloured Formatting for Columns	1.3.1
□	pkgconfig	Private Configuration for 'R' Packages	2.0.2
□	prettyunits	Pretty, Human Readable Formatting of Quantities	1.0.2
□	progress	Terminal Progress Bars	1.2.0
□	R6	Encapsulated Classes with Reference Semantics	2.4.0
□	Rcpp	Seamless R and C++ Integration	1.0.1
☑	readxl	Read Excel Files	1.3.1
□	rematch	Match Regular Expressions with a Nicer 'API'	1.0.1
□	rlang	Functions for Base Types and Core R and 'Tidyverse'	0.3.1

```
rying back to using insecure URL for this mirror.

To learn more and/or disable this warning message see the "Use secure download method fo
r HTTP" option in Tools -> Global Options -> Packages.
> library(readxl)
> twitterdata <- read_excel("C:/Users/GR SOFT/Desktop/twi/twitterdata.xlsx")
> View(twitterdata)
>
```

Figure 5.1.View Twitter Dataset



```
summary(cars)
summary(trees)
summary(kappa.default(cars))
summary(twitterdata$sentiment)
summary(twitterdata$comments)
runif(300)
```

```
Data
twitterdata    1038 obs. of 5 variables
Values
shuffle_index  int [1:1038] 228 693 855 502 703 786 144 297 17 443…
```

```
> hist(twitterdata$sentiment ,col="red",lwd="3")
> summary(cars)
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
> summary(trees)
     Girth           Height       Volume    
 Min.   : 8.30   Min.   :63   Min.   :10.20  
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40  
 Median :12.90   Median :76   Median :24.20  
 Mean   :13.25   Mean   :76   Mean   :30.17  
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30  
 Max.   :20.60   Max.   :87   Max.   :77.00  
>
```

Figure 5.2 Twitter Dataset Cart Algorithm

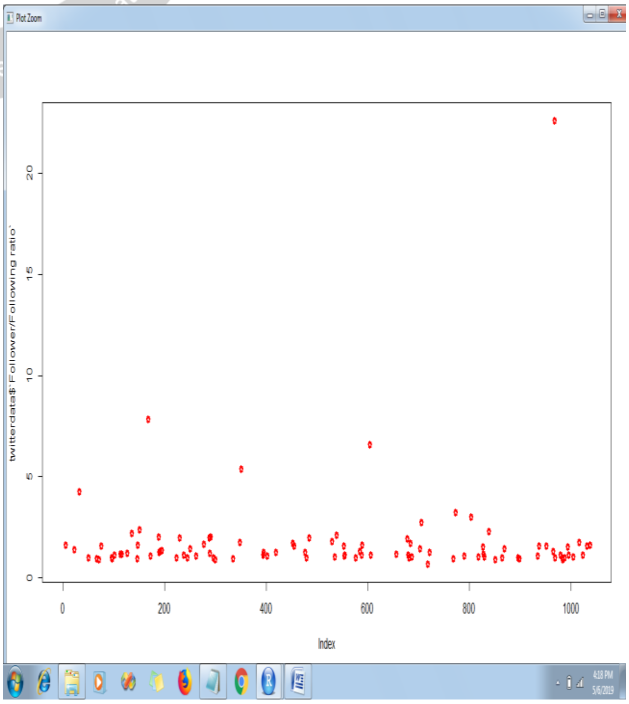
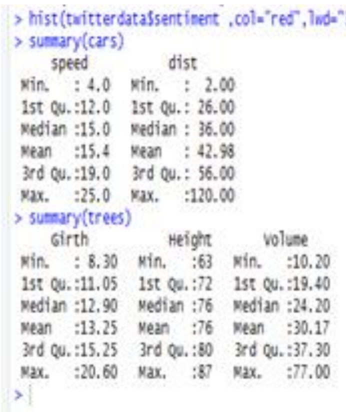


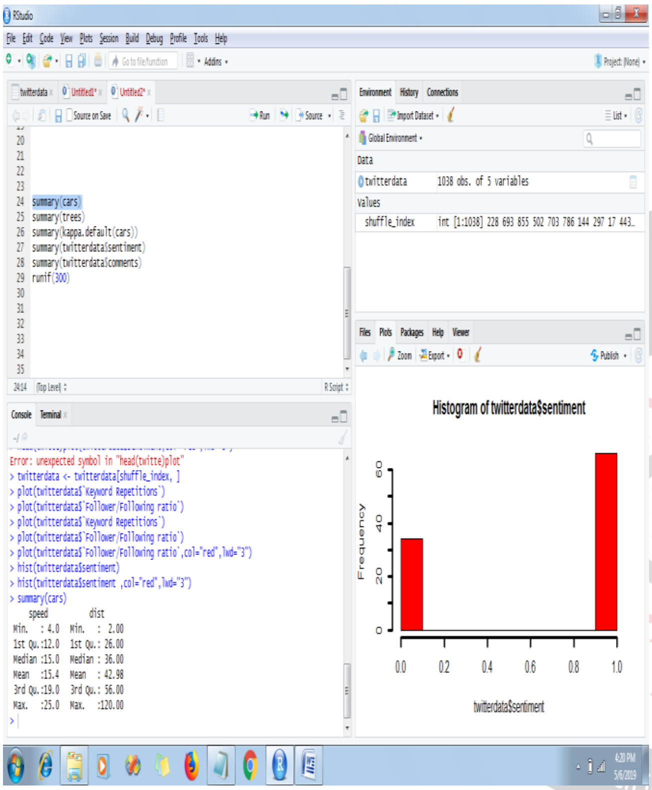
Figure 5.4.Twitter Dataset CART and Trees Algorithm

DOI : 10.35291/2454-9150.2019.0450



```
> hist(twitterdata$sentiment ,col="red",lwd="3")
> summary(cars)
     speed          dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
> summary(trees)
     Girth           Height       Volume
 Min.   : 8.30   Min.   :63   Min.   :10.20
 1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
 Median :12.90   Median :76   Median :24.20
 Mean   :13.25   Mean   :76   Mean   :30.17
 3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
 Max.   :20.60   Max.   :87   Max.   :77.00
> |
```

Figure 5.3 Twitter Dataset Plot View



```
summary(cars)
summary(trees)
summary(kappa.default(cars))
summary(twitterdata$sentiment)
summary(twitterdata$comments)
runif(300)
```

twitterdata 1038 obs. of 5 variables

shuffle_index int [1:1038] 228 693 855 502 703 786 144 297 17 443…

```
Error: unexpected symbol in "head(twitte)plot"
> twitterdata <- twitterdata[shuffle_index, ]
> plot(twitterdata$`Keyword Repetitions`)
> plot(twitterdata$`Follower/Following ratio`)
> plot(twitterdata$`Keyword Repetitions`)
> plot(twitterdata$`Follower/Following ratio`)
> plot(twitterdata$`Follower/Following ratio`,col="red",lwd="3")
> hist(twitterdata$sentiment)
> hist(twitterdata$sentiment ,col="red",lwd="3")
> summary(cars)
     speed          dist
 Min.   : 4.0   Min.   :  2.00
 1st Qu.:12.0   1st Qu.: 26.00
 Median :15.0   Median : 36.00
 Mean   :15.4   Mean   : 42.98
 3rd Qu.:19.0   3rd Qu.: 56.00
 Max.   :25.0   Max.   :120.00
> |
```

Figure 5.4.Twitter Dataset Trees Algorithm

VII. CONCLUSION

With the growing social network, it is challenging to analyze its large data using existing data mining tools. An experiments to do Sentiment Analysis on retrieved 'Twitter" data from Twitter that the number of people have given positive and negative opinions on the scheme CART and Trees. With this, it is advisable to conclude R Statistical Tool is sufficiently used for the analysis of Data mining. The decision tree and CART algorithms are used to analyze twitter dataset. By this analysis the mean, median and the execution second is been analyzed with the imported data. Decision tree algorithm gives the more accurate result when compared to the other algorithm. This algorithm will compute more data set with the less execution time. It is good at its performance and accuracy. C4.5 algorithm performs well when it compared to the CART algorithm

REFERENCES

[1] WalaaMedhat et al, "Sentiment analysis algorithms and applications: A survey", Ain Shams Engineering Journal (2014) 5, 1093-1113.

[2] Richa Sharma et al, "Opinion Mining of Movie Reviews at Document Level", International Journal on Information Theory (IJIT), Vol.3, No.3, July 2014.

[3] Rodrigo Moraes et al, "Document-level sentiment classification: An empirical comparison between SVM and ANN", Expert Systems with Applications 2012.

[4] Parminder Bhatia et al, "Better document-level sentiment analysis from RST Discourse Parsing".

[5] ZhaopengTu et al, "Identifying High-Impact Sub-Structures for Convolution Kernels in Document-level Sentiment Classification", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. July 2012.

[6] Matthew Hurst and Kamal Nigam, "Retrieving Topical Sentiments from Online Document Collections".

[7] Wei-Hai Lin et al, "Which Side are You on? Identifying Perspectives at the Document and Sentence Levels", In Proceedings of the Tenth Conference on Natural Language Learning (CoNLL'06).

[8] AinurYessenalina et al, "Multi-level Structured Models for Document-level Sentiment Classification", proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.

[9] Vivek Kumar Singh et al, "Computing Sentiment Polarity of Text at Document and Aspect Levels", Ecti Transaction on Computer and Information Technology Vol.8, No.1 May 2014.

[10] B. Pang, L. Lee, and S. Vaithyanathan, (2002), "Thumbs up? Sentiment classification using machine learning techniques" In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP).

[11] B. Liu, "Sentiment analysis and opinion mining," Proceedings of 5th Text Analytics Summit, Boston, June 2009.

[12] Cristianini, N., &Shawe-Taylor, J. (2000). An introduction to support vector machines and other Kernel-Based learning methods 1st ed. Cambridge University Press.

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

[14] Bing Liu, (2012), "Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers".

[15] Ronen Feldman. 2013. Techniques and applications for sentiment analysis.Communications of the ACM,56(4):82-89.

[16] Jeffrey Strickland, " Data analytics using open-source tools" , First edition (July 1, 2016), Lulu.com. [17] Guangxing Wang,Qihao Weng, Remote Sensing of Natural Resources, 1st edition CRC press book.

DOI : 10.35291/2454-9150.2019.0450