ISSN: 0974-5823 Vol. 7 No. 1 January, 2022

International Journal of Mechanical Engineering

Efficient Techniques for Analysing Real-Time Text Data Using Sentiment Analysis

Vishal C
Assistant Professor, RV College of Engineering, Bengaluru, India.

B.H.Chandrashekar
Associate Professor, RV College of Engineering, Bengaluru, India.

Abstract

In recent years, specialists from different orders like software engineering, insights and instruction have started to explore how opinion mining rules can be used to improve training and encourage training research. Extracting data from internet based life gives us a few utilization in different fields. The utilization of machine learning procedures has gotten progressively wide spread in business applications and scholarly research. Our research mainly focuses on various machine learning strategies and models which are reasonable for various issues, picking the correct strategy and fine-tuning its specific settings are critical errands that will straight forwardly affect the nature of the expectations.

Keywords:opinion miming, machine learning, sentiment analysis.

1 Introduction

Sentiment Analysis is the route toward changing rough data requested via preparing systems in accommodating information that could be used to take instructed decisions and answer research questions. Sentiment Analysis centre on creating techniques for investigating the interesting kinds of information that originates from an instructive setting. This information emerges from customary up close and personal (slate instructing) homeroom condition, instructive software's, and summarize/high stakes tests. The strategies can be used to assemble data that can help instructive architects to build up an academic reason for choices when planning or changing a domain's educational method. The use of information mining to the plan of the instructive framework is an iterative cycle of theory arrangement, testing, and refinement. It's surveyed that 80% of the world's data is unstructured and not dealt with in a for each portrayed way. The majority of this originates from content information, like messages, emails, chats, online networking, articles, and records. As of late, an immense number of people have been pulled into interpersonal interaction stages like Facebook, Twitter, and Instagram [1].Most use social locales to express their feelings, convictions or sentiments about things, spots or characters on social media. The increasing popularity of social media (such as online communities) has spawned a huge amount of information in society today.

Sentiment analysis is broadly utilized in different applications like Social media monitoring, Customer administration, Product investigation, Market research and examination and so on. By utilizing a concentrated opinion examination framework, the association can apply similar criteria to most of their information. This diminishes mistakes and improves information consistency. Opinion and recognition are crucial parts of human presence.

Content information can be comprehensively arranged into two principles types: facts and opinions. Actualities are target articulations about something [6]. Opinions are typically abstract articulations that depict individuals' slants, evaluations, and emotions toward a topic or theme.

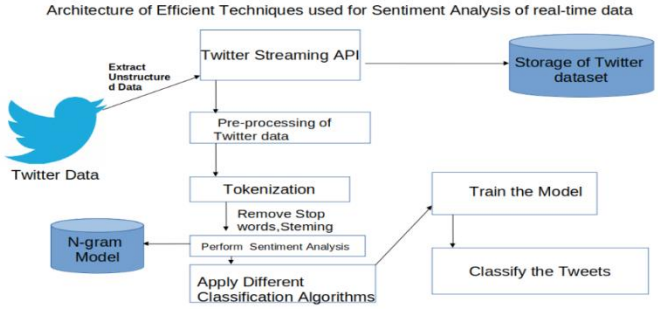

Figure 1. Architecture of Efficient Techniques used for Sentiment Analysis of real-time data

Figure 1 represents the architecture of using efficient techniques used for sentiment analysis of real-time data from twitter and processes in different forms of techniques.

2 Relatedworks

Slant investigation is the computerized procedure that utilizations of sentiment analysis to distinguish positive, negative and unbiased feelings from the content. Feeling investigation is generally utilized for getting bits of knowledge from web-based life remarks, study reactions, and item audits, and settling on information-driven choices. In reality, users can produce 2.5 quintillion bytes of information consistently; estimation investigation has become a key instrument for understanding that information.

2.1 Contextual features of Sentiment Analysis

DimitriosMichailidis et al.[11] discussed the real-time classification of data based upon the customer location. Tweets regarding the customer from different locations based on the present in the United States of America and various streams of real-time Twitter data have been classified into emotional contents like positive, negative and mixed terms. YanpingLv et al.[12] According to the author social networking data by the users plays a vital role in communication based upon the geolocation. Based on the trending events a visual analysis of the connection between network space and real physical space has been identified.

2.2 Opinion Based feature Extraction

Vladimir Gorodetsky et al.[13] The paper proposes another innovation bolstered by various novel calculations planned for philosophy focused change of heterogeneous conceivably poor organized learning information into homogeneous instructive double element space dependent on a conglomeration of the cosmology idea cases and their characteristic areas and resulting probabilistic reason outcome examination went for extraction progressively enlightening highlights. The proposed innovation is completely actualized and approved on a few contextual investigations. The proposed highlight extraction approach was completely actualized and approved utilizing a few applications. It was additionally utilized in the structure and execution of a cosmology based profiling and suggesting framework. Specifically, astute email colleague for approaching email arranging was prototyped.

2.3 Opinion Based Classification Technique

Krina Vasa et al. [14] Content grouping has become an essential procedure for taking care of and arranging content information. When all is said in done, Text order assumes a significant job in data extraction, content outline content recovery, restorative conclusion, newsgathering separating, spam sifting, and assumption examination. Content grouping is a definitive issue in information mining and AI. After investigating the papers, they confirm that there are such a large number of procedures in content characterization. Bolster vector machine, k-closest neighbour and nave Bayesian strategy are broadly utilized procedures in a content arrangement. The crosses breed approach of these systems likewise valuable in content characterization.

2.4 Data Pre-processing

Sivakumaret al.[15] Paper depicts an effective methodology for information pre-processing for mining based bioinformatics and web use mining information to accelerate the information arrangement process. This paper reviews the information pre-processing exercises like information cleaning, information decrease, and related calculations. It isn't just giving adaptability to information pre-processing, yet additionally lessens multifaceted nature and trouble in planning mining.Pre-processing improves the presentation of bioinformatics and web mining information. Information cleaning schedules can be utilized to fill in missing qualities, smooth boisterous information, recognizing exceptions, and the right information irregularities.

3 The proposed method

Text classification is a process of characterizing the text into two different phases: the preparation stage and the testing stage. Typically, the preparation stage incorporates making the named corpora dataset, pre-handling the preparation content, vectorization of the content, and preparing of the classifier. The testing stage incorporates the pre-handling of testing content, vectorization, and order of the testing content. Classification is the undertaking of learning an objective capacity f that maps each ascribes set X to one of the predefined class marks y. Order forecast includes two levels: classifier model development and the use of the input to the classification model is a collection of records.

Figure 2 Classification Model Mapping of Input and Output values

Step1:

3.1 Extraction of twitter Data using Open Authentication

Open Authentication (OAuth) is an open standard for validation, received by Twitter to give access to ensured data. Passwords are profoundly defenseless against robbery and OAuth gives a more secure option in contrast to conventional verification approaches utilizing a three-way handshake. It additionally improves the certainty of the client in the application as the client's secret key for his Twitter account is never imparted to outsider applications. The confirmation of API demands on Twitter is completed utilizing OAuth.

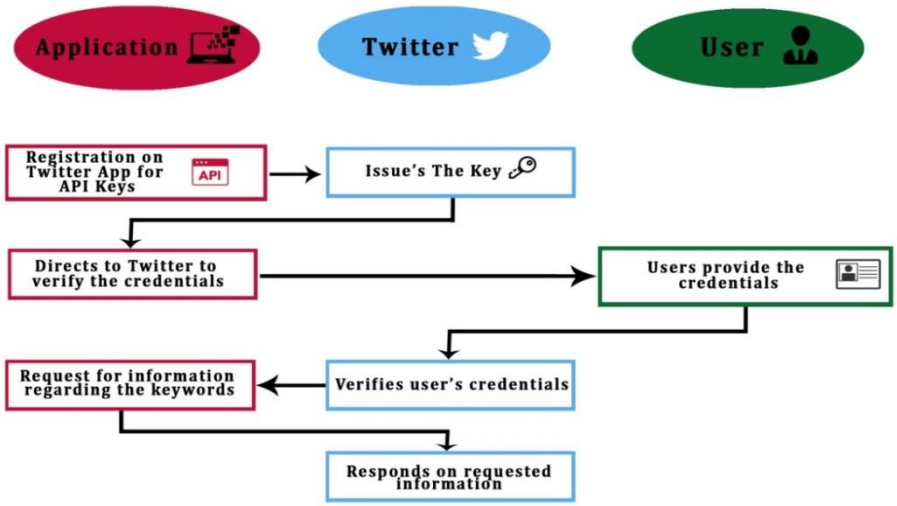

Figure 3 Open Authentication of Twitter data

One of the primary approaches was to order each tweet through the two classifiers and afterward figure the normal extremity, yet in the wake of estimating the precision and speed, the outcomes were not ideal the classification of data for any better and the speed diminished fundamentally a new outcome is expected to make a framework quick enough to be "continuous" and furthermore expanded inexactness. The proposed models are used for training and testing the text data. By implementing the algorithm the behaviour model can be determined.

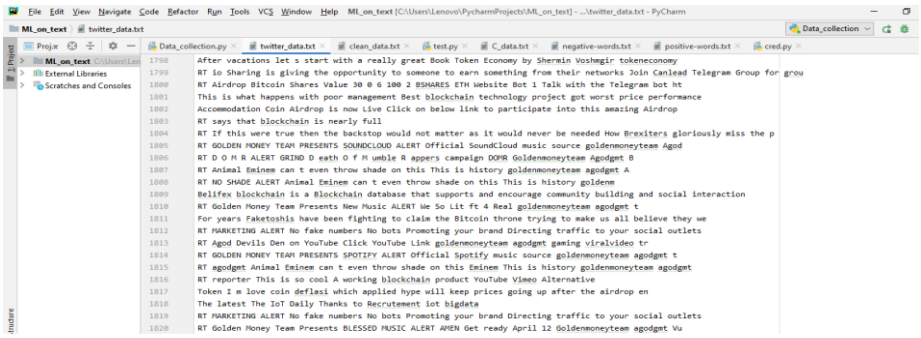

Figure 4 Collections of Tweets from Twitter Application

Step 2

3.2 Term Frequency and Inverse Document Scores

The document term matrix (DTM) is the collection of positive and negative terms. The rows represent the sentences in the document in the collections and columns represent the main terms of the documents and it is known as term frequencies. 'tm' built-in function package to create DTM.The Term Document Frequency (TDF) is processed for a lot of given survey archives x and a lot of terms a. The term recurrence is indicated by freq(x, a), speaking to the number of events of term an in the archive x. The Term Recurrence network TF(x, a) measures term a relationship as to the given record x and has an estimation of zero on report term for non-event or a number generally. The number can be set as TF(x, a) = 1, when term a shows up in the record x or when a relative term recurrence is utilized.

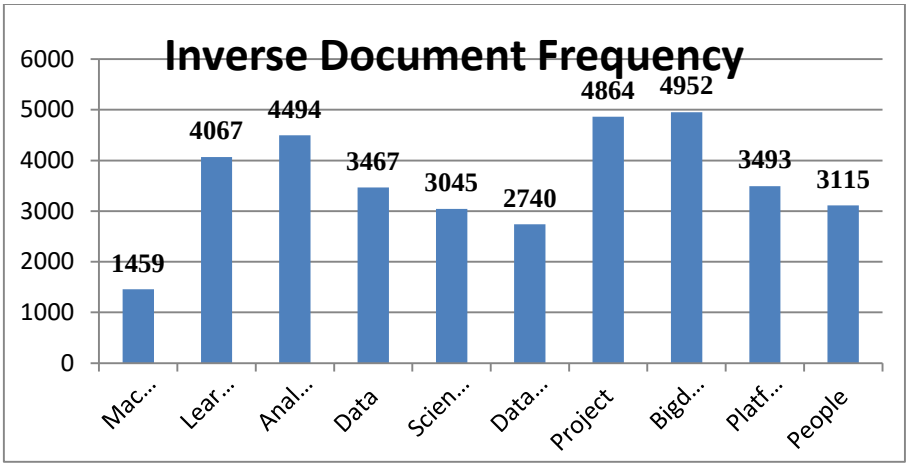

Figure 6 Graphical Representations of Term Frequency Scores

Step3:

3.3Proposed method for text classification model by using Navies Bayes classifier

The Naive Bayes Classifier strategy relies upon the indicated Bayesian theory and is particularly fit when the dimensionality of the wellsprings of data is high. Despite its straightforwardness, Naive Bayes can regularly defeat progressively propelled portrayal rules [59].

$$P(t|I) = \frac{P(I|t)\,P(t)}{P(I)}$$

Classification of terms used in Navies Bayes classifier which is described in the below process.

- P (t): the prospect of assumptions j being true. This is known as the preliminary prospect of j.
- P (I): the prospect of the data (regardless of the assumptions). This is known as the preliminary prospect.
- P (t|I): the prospect of assumptions j given the information I. This is known as a posterior prospect.
- P (I|t): the prospect of information I given that the prospect is j was true. This is known as a posterior prospect.

Step 4:

Applying Multivariate Bernoulli Naive Bayes

In the Multivariate Bernoulli occasion model, highlights are self-administering booleans (parallel factors) outlining out data sources. Like the multinomial model, this model is noticeable for portrayal depiction assignments, where twofold term occasion (for instance a word happens in a record or not) highlights are utilized instead of term frequencies (for instance rehash of a word in the report).

$$f(X) = \begin{cases} p \text{ if } X = 1 \\ q \text{ if } X = 0 \end{cases}$$

Where q = 1 - p and 0 < p < 1

$f(X)$- is considered to be a random number

p if X =1 and *q if X = 0* binary number

Step 4:

Algorithm: Multivariate Bernoulli Naive Bayes classifiers

Input: D: DataStream and N; Document

Output: Sentiment Classification Results

Step 1: Assume that the Number of Records in a document

Let N_{doc}= Number of records in R

N_c= Number of records from R in class c

Step 2: Calculate each records

logprior[c]←log $\frac{N_c}{N_{rec}}$

Step 3: Number of Records in Vocabulary (V)

V←vocabulary of R

Step 4: Appending each document with Bigdoc[C]

Bigdoc[C] ←append (r) for d ϵ R with class c

Step 5: Calculate P(w|c) terms

for each word w in V

Step 6: Occurrences of w in bigdoc[c] count (w,c)

Step 7: Calculating each count of scores of terms in the document

for each w Є V

then score[o] + = log condprob [w][c]

else score[o] + = (1- log condprob [w][c])

Step 5:

3.4 Proposed method for text classification model using Decision Tree classifier

A Decision tree is a guide of the potential consequences of the improvement of related choices. The entropy values are defined based upon the condition value which is shown I the equation 4.6 It empowers an individual or relationship to check potential exercises against one another subject to their costs, probabilities, and positive conditions [62].

Entropy= $-\sum \boldsymbol{p(X) log p(X)}$

P(X) - fraction of examples in a given class

Entropy= -p $\log_2$ – q $\log_2$ q

ALGORITHM: Entropy Decision Tree Model

Input: C, where C = set of classified instances

Output: Decision Tree

Require: C ≠ ∅, num_attributes> 0

Step1: procedure to Build Tree

Step2: repeat

maxGain←0

Split A←null

n←Entropy(Attributes)

Step3: Attributes for all attributes α in C do

gain←InformationGain(a,n)

Step4: If gain >maxGain then

maxGain←gain

Step5: **Leaf Node** Split A←a

end ifend for

Step6: Partition(C, splitA)

#Tuning Hyper Parameters for Entropy Decision Tree

clf = Decision Tree Classifier(criterion="entropy", max_depth=3,splitter="best")

Overall Analysis of classification algorithms

As indicated by this examination the best classifier dependent on the TP Rate, the FP rate, the Precision, the Recall, and the F-Measure is Bayes Net as it has the most noteworthy rates. As indicated by the ROC Area the Navies Byes (Multinomial Naive Bayes) has the most astounding rate.

4 Results

As indicated by this examination the best classifier dependent on the TP Rate, the FP rate, the Precision, the Recall, and the F-Measure is Bayes Net as it has the most noteworthy rates. As indicated by the ROC Area the Navies Byes (Multinomial Naive Bayes) has the most astounding rate. The comparative study of each model is classified and described in table 2

Table 2 Comparative Analysis of Classification Models

	Accuracy	**Sensitivity**	**Specificity**	**Precision**	**F1-score**
Navies Bayes	77	77.06	81.25	76.08	74.06
Navies Byes **(**Multinomial Naive Bayes**)**	83.87	83.09	66.67	87.02	82.07
Support Vector Machine(SVM)	77.47	75.00	74.00	82.00	70.04
Decision Tree (Gini Index)	83.76	60.00	80.00	87.01	82.06
J48 Decision Tree (Entropy)	74.29	70.00	80.00	82.35	75.68
Random Forest	83.87	83.09	83.09	87.02	82.07

5.2 Literature Review Comparative Analysis results

		Literature Data				Present Work			
SI NO.	Name of the algorithm	Author and Year	Datasets	Parameter used for analyzing the algorithms	Accuracy scores	Datasets	Parameter used for analyzing the algorithms	Accuracy scores	Comparison Results
1	Navies Bayes	Mokhairi makhtar 2017	Education results	10-fold cross validation as a test method	70.55	Technologies in Education	10-fold cross validation	77	Present method approaches better accuracy level
2	Multinomial Naive Bayes	Muhammad Abbas 2019	Movie reviews	Bow model and TF - IDF module has been incorporated with the algorithm	84	Technologies in Education	Aalpha=1.0, class_ prior=none, fit_ prior=True 10-fold cross validation	83.87	Present method performs better than the proposed method of 1% accuracy level
3	Support Vector Machine(SVM)	Bhumika M. Jadav 2016	Twitter datasets on news, product	hyper parameter value of RBF kernel	76.92	Technologies in Education	C=1.0, kernel='linear', degree=3, gamma='auto' and 10-fold cross validation as a test method	77.47	Present method approaches better accuracy level
4	Decision Tree	A. Suresh 2016	RatingSystem.com	LVQ type learning models	75	Technologies in Education	(criterion="entropy", max_depth=3,splitter="best" 10-fold cross validation	83.76	Present method approaches better accuracy level
5	J48	Josip Mesarić 2016	Education Data	The average rate of classification was calculated for two models	73.9%	Technologies in Education	C=1.0, kernel='linear', degree=3, gamma='auto' 10-fold cross validation	74.29	Present method approaches better accuracy level
6	Random forest	Palak Baid 2017	Movie Reviews	10 fold cross validation	78.65		(criterion="entropy", max_depth=3,splitter="best" And 10-fold cross validation	83.87	Present method approaches better accuracy level

The statistical classification (confusion matrix by positive and negative terms) of the Decision Tree (Gini Index) is predicted by the accuracy of 83.87, which is displayed in table 5.8.

Sensitivity, Specificity, and Accuracy are the terms that are most commonly associated with a Binary portrayal test and quantifiable measure the show of the test. In a joint gathering, we seclude a given instructive accumulation into two characterizations dependent on whether they have standard properties or not by recognizing their significance and in a twofold plan test, Sensitivity shows the result of how well the test predicts one classification and Specificity gauges how well the test predicts the other classification.

5 Conclusions

The information from the twitter data results in an attractive source of information for opinion and sentiment analysis. The prediction of tweets over the top trends helps the user to know the opinion of other user's human behaviour. Our classifier will make use of classification techniques, aspect extraction and supervised machine learning algorithms. Navies Byes (Multinomial Naive Bayes) algorithm has been found as one the best strategy to bring out continuous things set just as intriguing guidelines with regards to discovering the conduct of the understudies towards the learning style. It is discovered that the understudies are moving from customary or sound-related sort of learning towards visual learning. Through this sentiment analysis, any education organization can adopt these kinds of technologies and build a gap between academia and the industry. They are likewise equipped for characterizing the guidelines as per the qualities present in the survey. The yield as Sensitivity, Specificity, and Precision esteems additionally clarifies the skill of the calculation.to build up a procedure to join diverse supposition examination strategies that yielded an exact slant investigation arrangement in a continuous domain utilizing Twitter as the wellspring of substance. The proposed methodology addresses the issue of arranging tweets into conclusion classes when marks are scarce.Micro blogging on

Twitter plays a vital role in the executions of user's opinion on a particular topic and differ too much on relying upon the kind of information that an assessment examination framework needs to be ordered and classified.

6 References

[1] Kawaljeet Kaur Kapoor,Kuttimani Tamilmani,Nripendra, P. Rana," Advances in Social Media Research: Past, Present and Future", Information Systems Frontiers, Volume 20, Issue 3, pp 531–558, 2018.

[2] Abdullah Alsaeedi, Mohammad ZubairKhan,"A Study on Sentiment Analysis Techniques of Twitter Data", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 2, pp 361, 374, 2019.

[3] Peter Fernandez, "Online Social Networking Sites and Privacy: Revisiting Ethical Considerations for a New Generation of Technology", Library Philosophy and Practice, ISSN 1522-0222, pp 1-9, 2015.

[4] Cha Oyza, Agwu Edwin, "Effectiveness of Social Media Networks as a Strategic Tool for Organizational Marketing Management ", Journal of internet and banking and commerce, ISSN: 1204-5357,pp 2-9. 2015.

[5] TahuraShaikh and DeepaDeshpande, "A Review on Opinion Mining and Sentiment Analysis", Proceedings on National Conference on Recent Trends in Computer Science and Information Technology NCRTCSIT 2016(2): pp 6-9, 2016

[6] DibyenduMondal,"Study of Significance Tests with respect to Sentiment Analysis", RnD Project, Department of Computer Science and Engineering Indian Institute of Technology, Bombay, 2019.

[7] ValentinSchöndienst, Hanna Krasnova, "Micro-Blogging Adoption in the enterpriseAnalysis",10th International Conference on Wirtschaftsinformatik, , pp 1-17, F 2011.

[8] AlessiaD'Andrea, "Approaches, Tools and Applications for Sentiment Analysis Implementation" ,International Journal of Computer Applications Volume 125 – No.3, ,pp 0975 – 8887, 2015.

[9] Abibollah. NaderiRohani. Abdullah, H. TengkuAizan, Jamaluddin. Sharir, "Intelligence and academic achievement: an investigation of gender differences", Life Science Journal, Vol 7, No 1, pp 83-87. 2010.

[10] Sameer Hinduja, Justin W. Patchin, "Personal information of adolescents on the Internet A quantitative content analysis of MySpace", Journal of Adolescence, pp 125–146, 2008.

[11] DimitriosMichailidisi, TomoakiOhtsuki, "Real-Time Location-Based sentiment Analysis on Twitter the AirSentSystSETN '18, July 9–15, , Rio Patras, Greece, 2018.

[12] Yanping Li, Xiao, Dazhen Lin, Donglin Cao, "Public Opinion Analysis Based on Geographical Location", 8th International Congress on Image and Signal processing ,2015.

[13] Vladimir Gorodetsky, Vladimir Samoylov, "Feature Extraction for Machine Learning: Logic–Probabilistic Approach", JMLR: Workshop and Conference Proceedings 10: the Fourth Workshop on Feature Selection in Data Mining,PMLR 10:55-65, pp 55-65, 2018.

[14] Krina Vasa, "Text Classification through Statistical and Machine Learning Methods: A Survey", International Journal of Engineering Development and Research, Volume 4, Issue 2, ISSN: 2321-9939, pp 655-658, 2016

[15] A.Sivakumar and 2R.Gunasundari, "A Survey on Data Pre-processing Techniques for Bioinformatics and Web Usage Mining", International Journal of Pure and Applied Mathematics, Volume 117 No. ISSN: 1314-3395 pp 785-794, 2017.

[16] Jadhav, Snehal. "An internet based interactive Embedded data Acquisition system for real time application." International Journal of Engineering Research & Technology. International Journal of Electronics, Communication & Instrumentation Engineering Research and Development (IJECIERD) 4.3 (2014): 15 22.

[17] Hymavathi, Ch, And Y. Rama Mohan. "An Adaptive Scheme Over Text Streams For Real-Time Monitoring." International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) 8.4, Oct 2018, 1-6

[18] Jayaram, B., et al. "A Survey On Social Media Data Analytics And Cloud Computing Tools." International Journal of Mechanical and Production Engineering Research and Development, 8 (3), 243 254 (2018).

[19] Caballero, Arlene R., Jasmin D. Niguidula, And Jonathan M. Caballero. "Twitter Feeds Sentiment Analysis And Visualization." International Journal of Educational Science and Research (IJESR) 7 (2017): 31-40.

[20] Khan, Mudassir, Aadarsh Malviya, And Surya Kant Yadav. "Big Data Approach Of Sentiment Analysis Of Twitter Data Using K-Mean Clustering Approach." International Journal of Mechanical and Production Engineering Research and Development (IJMPERD) 10.3, Jun 2020, 6127-6134

[21] Deepa, S., and R. Umarani. "Steganalysis on images based on the classification of image feature sets using SVM classifier." International Journal of Computer Science and Engineering (IJCSE) 5.5 (2016): 15-24.