ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 7| Issue: 10| October 2021|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2021: 8.047 || ISI Value: 1.188

DISCERNMENT OF CYBERBULLYING APPROACH

Shruti Aggarwal[1], Himanshu Sharma[2], Sanjana Mann[3], Mehul Gupta[4]

[1]*Jagan Institute of Management Studies, Rohini, Delhi, India*

[2]*Jagan Institute of Management Studies, Rohini, Delhi, India*

[3]*Jagan Institute of Management Studies, Rohini, Delhi, India*

[4]*Jagan Institute of Management Studies, Rohini, Delhi, India*

Article DOI: https://doi.org/10.36713/epra8697
DOI No: 10.36713/epra8697

ABSTRACT

People in the 21st century are being raised in an Internet enabled world where social media has become an integral part of people's daily routine, with communication just a click away. As per the latest survey, the number of individuals using social media is over 39.6 million worldwide, with the average user having 8.7 accounts on various networking sites. Social media provides an opportunity to connect with people and share data in the form of posts, text etc, with this package of pros and yet various individuals are trying to misuse it by spreading hatred towards a group, individuals, a topic or an activity. Due to which cyberbullying has come into play, affecting the psychological state of the person. Where prevention is much needed, for which many researchers have come together and established many such technologies and programs for automatically detecting the events of cyber bullying on social media and preventing them by analysing the pattern of the posted comments or images. Thus the purpose of this research is to track and monitor the threats using supervised machine learning and mining.

1. INTRODUCTION

Over the past decade or two, social media has proliferated at an unprecedented rate which has changed the way people communicate among themselves by sharing text, images and videos leaving behind the traditional ways.

In the year 2020, a major boom was noticed on the social networking sites as the world was hit by a pandemic namely COVID-19, which bounded people to stay inside and away from physical interaction, so social media became a very easier means to interact with their famed over distances and also entertain them whilst all the negativity.

As the facts suggests, in 2020, around 3.97 billion people were using social media in the world, and this was an increase of 10.9% from 3.48 billion in 2019.

Around 2021, there are 37.8 million social networking sites users which equates to around of population. Also on an average 2.5 hours is spent daily per person on social media platforms which also led to negative impacts on ones life. One of the major concerns of over usage of social media is cyberbullying.

Cyberbullying can be defined as an aggressive or intentionally carried out harassment by a group or an individual through digital means repeatedly against a sufferer who is unable to defend themselves. This type of bullying includes threats, abusive or sexual remarks, rumours and hate speech.

According to the survey, 40% of teenagers and over 37% of adults across 32 countries were involved in the act of cyber bullying. also combing the facts, it was analysed that out of 38% that were involved, 19% were identified as the "aim," 21% as a "witness" and 1% each as both "supporter" and "bully."

After understanding the mindset and their bullying behaviour, it was discovered that some felt social pressure to act and some of them even regret their actions.

Victim of cyber bullying are more likely to experience depression, anxiety, suicidal ideation and even sometimes leads to suicide.

In order to overcome this situation of cyber bullying many techniques are being used. This paper would help us understand about the techniques and algorithms like SVM (Support Vector Machine) and TF-IDF (Term Frequency – Inverse Document

Frequency) which are used by various social media sites in particular Twitter.

Support Vector Machine is one of the most popular techniques used for classification and regression in Machine Learning.Term Frequency – Inverse Document Frequency is a term used in retrieval of information. It determines the frequency of words in a document and its inverse document frequency.

This paper consists of Literature Review which enlightens us about various machine learning algorithms, comparison with the best ML classifiers followed by their result and discussions.

2. LITERATURE REVIEW

This paper includes research work done by various researchers on techniques like SVM and TF-IDF and comparing them with others.

Yin et al.[1], used a supervised learning approach to detect harassment from three different social websites. Dataset along with the content features, sentiment features and contextual features of documents were used from Kongregate , Slashdot and Myspace. They used a lib SVM with the linear kernel as a classification tool. It was concluded that TFIDF was better than n-gram and Foul Language with higher weighing performance.

M. Dadvar and F.de Jong[2]., analyzed the gender approach within the cyber bullying detection problem, applied to the social network MySpace. Authors investigated the content of the posts written by the users but regardless of user's profile information. They used an SVM model to train a specific gender text classifier. The dataset consists of about 381,000 posts. The results obtained by the gender based approach improved the baseline by 39% in precision, 6% in recall, and 15% in F-measure

Chavan and Shylaja [3] also produced a score signifying other users the probability that a statement could be offensive. The accuracy was increased by 4% using a dataset from Kaggle10 by integrating the outcomes of Support Vector Machine classifiers.

Cynthia Van HeeID, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Ve ´roniqueHoste [4] has built a classifier which detects indication of cyberbullying on social media platform discriminating roles in the annotation scheme which includes victim, bully, bystanders-defendant and bystanders-assistant using Linear support vector machine as a classifier. They demonstrated the method which can be used for languages easily. The experiments were performed in English and Dutch datasets.

Bhatia et. al. in 2013, 2015 [19]proposed secure group communication techniques using steganography for communication of secret messages on the internet. To maintain the security of secret messages Bhatia in 2014 proposed image steganography method using spread spectrum approach. In the proposed technique author uses the properties of orthogonal image planes and secret message is modulated using pixels of one image plane of cover image. The modulated message is then replaced by the pixels of another image plane.

Bhatia, 2017 [20] proposed a secret message hiding technique, in this technique author divided the image into RGB planes. Each plane is further sectioned into 8*8 pixel blocks and secret message characters are embedded in pixels corresponding to position of 8-Rooks in 8*8 chessboard.

Bhatia in 2019 [21], proposed a message hiding technique, in this technique author used solutions of Knight tour and 8-Queen's problem in an 8*8 chessboard. The proposed technique applied solutions of moving knight tour and of placing 8-Queen's in non-attacking manner in 8*8 chessboard to select pixels for embedding secret message bits.

M. A. Al-garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A.Khattak, and A. Gani, [5], different algorithms were compared for detection of cyberbullying on social media which consists of SVM(Support Vector Machine), NB(naïve Bayes),RF(random forest),DT(decision tree),KNN(k nearest neighbour),LR(logistic regression),ARM(association-rule mining),RB(rule based algorithm).Out of all this the SVM algorithm is the best based on factors like accuracy, precision recall .The limitation of this paper is the unexplored deep learning architecture.

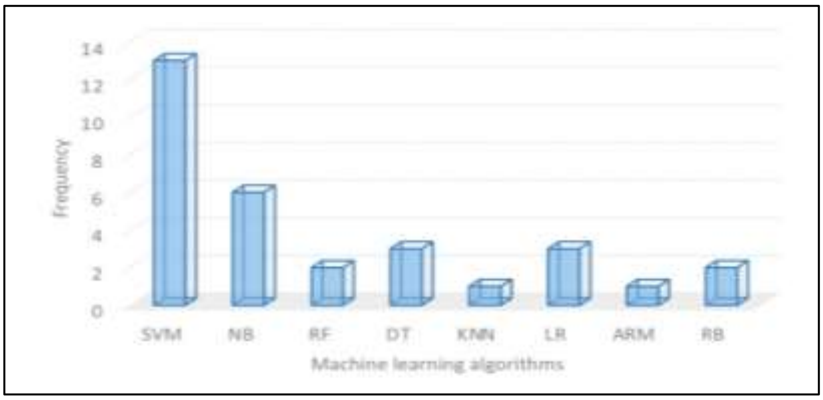


Figure 1[5]. Different Machine learning algorithms applied in cyber bullying detection.

3. METHODOLOGY

In this study, selection of a social networking site played a vital role for detecting cyber bullying. Thus after a regress research Twitter was selected as it generates loads of data everyday. As dataset was publicly available, a large amount of cyberbullying was noticed.

This dataset consists of multilingual unstructured content which was needed to be cleaned for higher accuracy rate. All data for this research was collected from Internet Archive [6], is an American open source digital library of websites, software app/games, videos, millions of books etc.

3.1. Research

During this research, a problem was identified which was to find a suitable technological approach that will help in detection of cyber bullying on social media. The approach explored, is a system capable of detecting and reporting incidents of cyberbullying on social media platforms.

The research was carried out using two machine learning approaches:

3.1.1. SVM: Support Vector Machine is one the most convenient and efficient classification algorithms. It uses supervised learning approach for classification. In SVM, 70% of its data is used for training purpose and rest 30% for testing [9]. Its main goal is to separate the hyperplane in such a way that it maximizes the margin of training data.

Where, fit() is used to fit the model into the train set and predict() is used to perform prediction on the test [9] . In simpler terms, the classifier is first trained with labelled data before it can be used to classify the data to evaluate accuracy, recall, precision. Once the classifier is trained with the labelled data, the input data is given to this classifier to separate it into positive and negative instances of bullying.

For instance [7]:

- "This girl is moron, I don't like it" which only includes a profanity so is considered not cyberbullying.
- "You are a bitch" in these case this is a cyberbullying and profanity along with a second person(you) or third person (She, He, They, It) or a person's name.

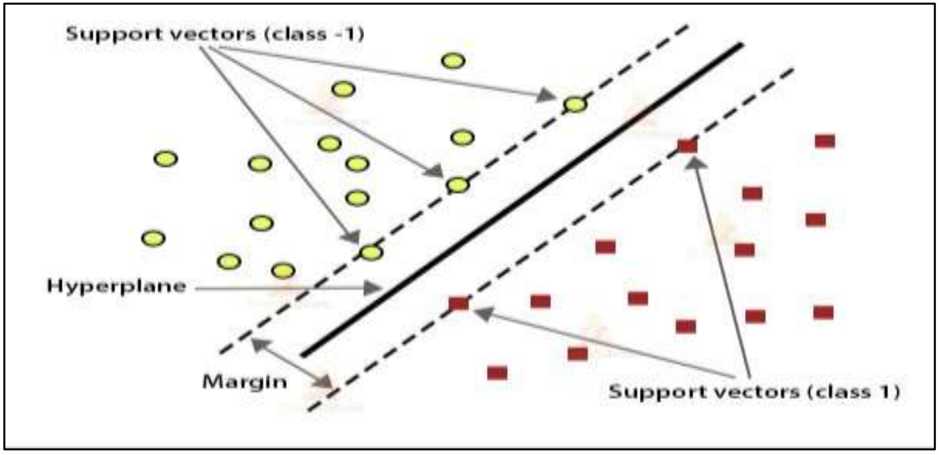


Figure 2[8]. Support Vector Machine

3.1.2. TF-IDF: Term Frequency, Inverse Document Frequency is a technique which is used to discover meaning of sentences consisting of words and remove the in capabilities of Bag of Words technique which is good for text classification or for helping a machine read words in numbers.

For a term t in document d, the weight Wt,d of term t in document d is given by:

$W_{t,d} = TF_{t,d} \log (N/DF_t)$ (1)

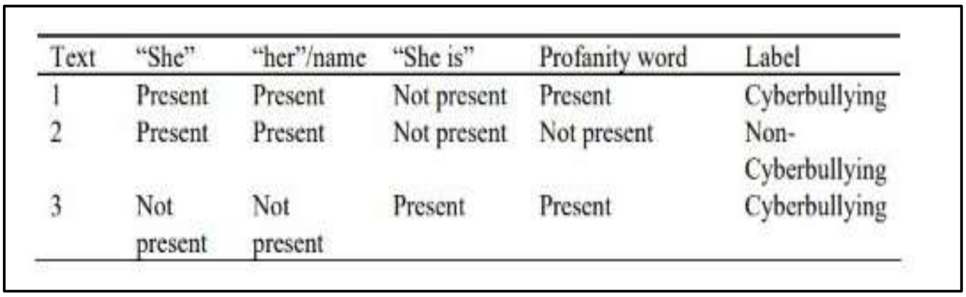
Where: $TF_{t,d}$ is the number of occurrences of t in document d.

DF_t here is no of documents that contain the term t.

N here is the total no of documents present in corpus.[10]

By using TF-IDF we can measure the importance of words in a document and Common words such as "is", "am" which do not affect the results due to IDF.

Profanity along with Pronoun- Most of the cyberbullying content contains profanity.Profanity alone cannot detect cyberbullying. Some studies, hypothesizes that cyberbullying text contains a swear /insult word along with the second person or third person (She, He, They, It, you) or a someone's name. The same is explained below in the table.



Text	"She"	"her"/name	"She is"	Profanity word	Label
1	Present	Present	Not present	Present	Cyberbullying
2	Present	Present	Not present	Not present	Non-Cyberbullying
3	Not present	Not present	Present	Present	Cyberbullying

Figure 3[7]. Labelling The Text

3.2. Methods

The detection of cyberbullying activities begin with extracting dataset from social network, where Input is text conversation, collected from Internet Archive.

3.2.1. Data Collection: The performance of the classifier, in this case SVM depends on the quality and size of dataset, which can be gathered by crawling the twitter data using Twitter API(rtweet , Twitter4j,Twit) and Tweepy Library5[11]. We crawl the twitter data using abusive words references as the query.

The data collected should consists of 3 attributes user id, group and comment. User id is used for the identification of a user, group attribute is used to recognize groups and comment defines the user comment on various status/group.

Once the dataset has been prepared, it has to be split into texts which includes comments,chats, etc. and media(images,videos,audio,etc.).

3.2.2. Data Pre-Processing: The data pre-processing is an important phase as Twitter data is noisy, thus pre-processing has to be applied to improve the quality of the extracted data and this includes removing all emoticons, folksonomies, slangs and stop words as they are not required for our purpose.

The pre-processing step is done in the following: -

- **Lowering text**: This then takes the list of words that we got from tokenization and then change them to lowercase letters Like: 'YOU ARE BEAUTIFUL' is going to be 'you are beautiful'.
- **Stop words and encoding cleaning**: This is an essential part of the pre processing where we clean the text from those stop words and encoding characters like \n or \t which do not provide a meaningful information to the classifiers.

3.2.3. Data Labelling: After data cleaning, the dataset is divided into two categories: training set and testing set. Where each dataset is labelled as bullying or no bullying.

3.2.4. Word Frequency analysis: After cleaning the dataset in the above steps , tokens can be extracted from it. The process of extracting tokens in known as Tokenization **,** where we take the extracted data as sentences or paragraphs and then output the entered text as separated words, characters or sub words in the form of a list.

These words need to be converted to numerical vectors so that each dataset can be represented in the form of numerical data. The vectorization of features are done using TF-IDF score which is helpful in balancing the weight between most frequent or general words and less commonly used words.TF-IDF value shows the importance of a token to a document in the corpus.

Example of Calculation of TF-IDF value:

Suppose a video comments contain 80 words wherein the word great appears 4 times.

The Term Frequency (i.e., TF) for great then (4 / 80) = 0.05.

Again, suppose there are 100000 video comments in the dataset and the word great appears 1000 times in whole corpus Then, the Inverse Document Frequency (i.e IDF) is calculated as log(100000/1000) = 2. Thus, the TF-IDF value is calculated as: 0.05 * 2 = 0.1.

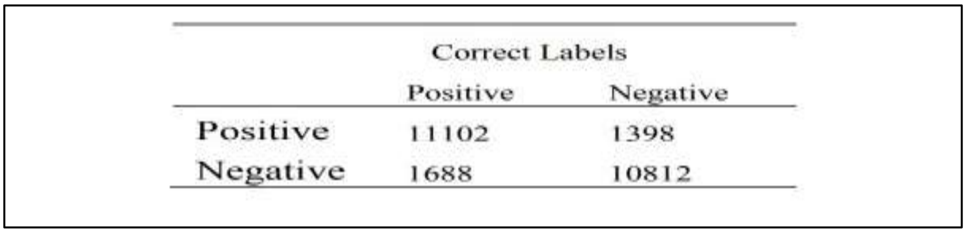
3.2.5. Classification

The numeric vectors can be given as input to the classification algorithm. Here SVM is used as an classification algorithm.

Support vector machine (SVM) algorithm: An SVM model is the representation of data as points in space mapped so that the examples of the separate categories are divided by a clear gap that is wide as possible. SVM plots all the numeric vectors in space and defines decision boundaries by hyperplanes. Support Vector Machine is a discriminative classifier formally defined by a separating hyper plane.

Here the given labelled training data uses the algorithm to give the optimal hyper plane which can classify new data This hyperplane separates the vectors in two categories such that, the distance from the of each category to the hyperplane is maximum 5. In addition to this SVM's can efficiently perform a non-linear classification, implicitly mapping their inputs into high dimensional feature space.

- It separates the training data set into two categories using a large hyperplane, that is in bullying context , positive and negative.
- After separating the training data, a matrix is generated known as confusion matrix and it shows the number of positive and negative words that are predicted right and number of positive and negative words that are predicted wrong.
- For each fold, prediction accuracy is determined on the basis of this confusion matrix and final accuracy is given by calculating the mean of all the individual accuracies of 10 folds. However, individual accuracy of a particular fold can be much higher than the mean of all accuracy.



	Correct Labels	
	Positive	Negative
Positive	11102	1398
Negative	1688	10812

Figure 4[13]. Confusion matrix for Support Vector Machine Classifier

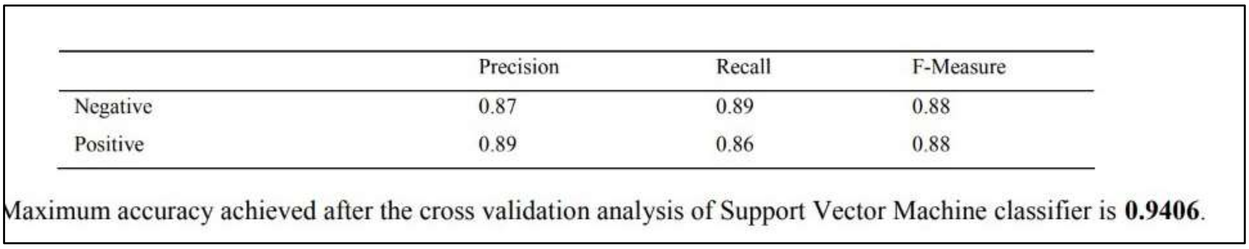
From this confusion matrix, different Performance evaluation parameter like precision, recall, F-measure and accuracy are calculated. The table of confusion matrix formation is shown in table 1.

Precision: It gives the exactness of the classifier. It is the ratio of number of correctly predicted positive reviews to the total number of reviews predicted as positive.

Recall: It measures the completeness of the classifier. It is the ratio of number of correctly predicted positive reviews to the actual number of positive reviews present in the corpus.

F-measure: It is the harmonic mean of precision and recall. F-measure can have best value as 1 and worst value as 0.

Accuracy: It is one of the most common performance evaluation parameter and it is calculated as the ratio of number of correctly predicted reviews to the number of total number of reviews present in the corpus.



	Precision	Recall	F-Measure
Negative	0.87	0.89	0.88
Positive	0.89	0.86	0.88

Maximum accuracy achieved after the cross validation analysis of Support Vector Machine classifier is **0.9406**.

Figure 5[12]. Evaluation parameters for Support Vector Machine classifier

3.3. Results

At this stage, researchers have carried out various experiments which indicates the results that are model achieves reasonable performance and could be usefully applied to monitor the heavy social problem of cyberbullying. It was observed that **tf-idf** is widely used for word frequency analysis and **svm** uses numeric vectors to define hyperplane and then calculates the precision of positive and negative aspects of a data.

An algorithm is presented by the researchers to analyse and track cyberbullies.

```
Begin {
Step 1: Pre-process the data (d)
Step 2: Divide the processed data into each
comment (c)
Step 3: Word Tokenization of each comment (tok)
Step 4: Repeat while (i <= d)
 Repeat (j <=tok)
 If (word == positive)
countPos = countPos++
else
countNeg = countNeg++
 end
 Polarity = countPos – countNeg
 end (step 4)
 }
```

The above mentioned algorithm[14], is used to measure the polarity of a particular cyber bully based on the number of positive and negative words used by him. it can be that cyberbully is frequent or not.

As per the algorithm, the polarity is measured for various users for different categories : political , religious , sports and terrorism. Based on this polarity , if polarity is positive the user is considered as non bully else bully.

4. CONCLUSION

Cyberbullying has many negative impacts on someone's life which includes depression, anxiety, anger, fear, trust issues, low self-esteem, exclusion from social events and sometimes suicidal behaviour too.This paper tries to address the issue of cyberbullying in media-based social network.

It is an appropriate definition of cyberbullying that incorporates both frequency of negativity and imbalance of power is applied in large- scale labelling, and is differentiated from cyber aggression. This proposed model will help cyber investigators and researchers pursuing the task of cyberbullying detection.

This research study focused on estimating the users behaviour's. A dataset consists of comments and replies by users on the Twitter were collected to perform descriptive analysis to analyse if a bully word was used in a positive or negative context and further identifying the most negative comments by a user in a particular group. The experiment was performed based on word-wise tokenization approach with the help of existing sentiment lexicons. The research work can be extended to analyze different Twitter groups or community pages to identify any unusual or offensive posts by the people against government agencies or others.

5. REFERENCES

1. *L. Hong and D. Yin, Z. Xue, "Detection of Harassment on Web 2.0," p. 8, 2019.*
2. *M. Dadvar and F.de Jong. 2012."Cyberbullying detection : a step toward a safer internet yard". ACM, New York, NY, USA, 121-126.*
3. *V. S. Chavan and Shylaja S S, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, Aug. 2015, pp. 2354–2358, doi: 10.1109/ICACCI.2015.7275970.*
4. *Cynthia Van HeeID, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, Ve ́roniqueHoste, "Automatic detection of cyberbullying in social media text", PLOS ONE, October 2018.*
5. *M. A. Al-garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A.Khattak, and A. Gani, „„Predicting cyberbullying on social media in the big data era using machine learning algorithms:Review Of Literature-And-Open-Challenges,""IEEEAccess,vol.7, pp. 70701–70718, 2019.*
6. *"Internet Archive Search: collection:twitterstream." https://archive.org/search.php?query=collection%3Atwitterstream&sort publicdate&page=2 (accessed Jul. 23, 2020).*
7. *Accurate Cyberbullying Detection and Prevention on Social Media by Andrea Pereraa, Pumudu Fernandob ,2021.*
8. *SVM in R for Data Classification using e1071 Package by https://techvidvan.com/tutorials/svm-in-r/.*
9. *Detecting Cyber Bullying On Twitter Using Machine Learning Techniques by Prajakta Ingle , Ramya Joshi ,Neha Kaulgud , Aarti Suryawanshi ,Meghana Lokhande , 2020.*
10. *http://casciolaw.attorneytestsite.com/learn-kaboat-amiibo/td-sequential-algorithm.html.*
11. *https://developer.twitter.com/en/solutions/academic-research/resources.*
12. *Comparative Study of Cyberbullying Detection using Different Machine Learning Algorithms by Rohini K R, Sreehari T Anil, Sreejith P M, Yedumohan P M , 2020 vol 4.*
13. *Classification of Sentimental Reviews Using Machine Learning Techniques by Abinash Tripathy,*, Ankit Agrawal, Santanu Kumar Rath,2015.*
14. *Development of Aggression Detection Technique in Social Media August 2019International Journal of Information Technology and Computer Science 11(5) DOI:10.5815/ijitcs.2019.05.05 Authors:Muhammad Asif.*
15. *https://www.researchgate.net/.*
16. *https://www.ieee.org/.*
17. *https://scholar.google.com/.*
18. *https://www.springer.com/in.*
19. *Bhatia, M. P. S., Bhatia, M. K., and Muttoo, S. K. ."Secure Group message transferring stegosystem", International journal of information security and privacy, IGI global (ISSN: 1930-1650), Vol. 9 no. 4, pp. 59-76, 2015. Bhatia, M. K, Muttoo, S. K. and Bhatia, M. P. S*

(2013). "Secure group communication with hidden group key", Information security journal: a global perspective, Taylor and Francis (ISSN:1939-3555 EISSN:1939-3547), Vol. 22 no.1, pp.21-34, 2013. Bhatia, M. P. S, Muttoo, S. K. and Bhatia, M. K (2014). "An Image Steganography Method Using Spread Spectrum Technique" in Springer sponsored International Conference on Soft Computing for Problem Solving (SocProS 2014) organized by NIT SILCHAR, Assam, India in 2014, pp:219-236.

20. *Bhatia, M. K.(2017), "8-Rooks Solutions for Image Steganography Technique", International journal of Next-Generation computing (ISSN: 2229-4678 (Print) and 0976-5034 (Online)), Vol. 8 no. 2, pp. 127-139, July 2017.*
21. *Bhatia, M. K.(2019), "Knight Tour for Image Steganography Technique", International Journal of Engineering and Advanced Technology (IJEAT)(ISSN: 2249 – 8958), Volume-9 Issue-1, pp. 1610-1613, October 2019.*