iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54217

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Machine Learning Algorithm for Toxic Comments Analysis

K Ramya Sri[1], Kommineni Madhavi[2]
[1, 2]*Asst Professor, Computer Science and Technology, G. Narayanamma Institute of Technology and Science, Hyderabad, India*

Abstract: ***Online comments that are toxic tend to drive other users away from a discussion because they are nasty, abusive, or irrational. The risk of online bullying and harassment limits people from expressing opposing views, which has an impact on the free exchange of ideas. Sites fail to properly promote discussions, which forces many communities to restrict or disable user comments. In order to analyze the toxicity as accurately as possible, this paper will carefully evaluate the prevalence of online harassment. In order to solve the text classification problem and determine the optimal machine learning algorithm based on our assessment metrics for the categorization of harmful comments, we will employ six machine learning algorithms and apply them to our data. We will work to accurately assess the toxicity to reduce its negative consequences, which will encourage organizations to take the necessary action.***

Keywords: Machine Learning, Toxic comments, Least Square Support Vector Machine, Singular Value Decomposition,

I. INTRODUCTION

With the use of just a smart phone and the internet, one person may now interact with another person anywhere in the world, which is one of the greatest advancements of the "Internet" of the century thanks to the growth of computer science technology. In the past, email was the only form of communication between people, and it was replete with spam. Classifying the emails as spam or not was a difficult task. As time went on, communication and data flow through the internet underwent a significant change, particularly following the advent of social media websites. With the development of social media, it is crucial to categorize the content into positive and negative words in order to stop any type of harm to society and to regulate people's antisocial behavior.

II. PROBLEM DEFINITION

A Web of Hate: Tackling Hateful Speech in Online Social Spaces proposes a way to distinguish scornful discourse that utilizes content created without help from anyone else recognizing disdainful networks as preparing information. This methodology sidesteps the costly explanation measure often needed to prepare catchphrase frameworks and performs well across a few set up stages, making improvements over present status of the craftsmanship. Provocation is a "highlight" of life online for some Americans, but it can bargain clients' protection, drive them to pick when and where to partake on the web, or even represent a danger to their actual wellbeing. This paper presents the Internet Argument Corpus (IAC), a collection of 390,704 posts in 11,800 conversations removed from 4forums.com. It examines the connection between talk marker pragmatics, understanding, emotionality, and mockery in the IAC corpus.

It finds that introverted conduct is more regrettable than other clients over the long run, and is exacerbated when local area input is excessively cruel. It also uncovers particular gatherings of clients with various degrees of reserved conduct that can change over the long haul.

This paper examines introverted conduct in three online conversation networks by dissecting clients who were prohibited from these networks. It is found that these clients move their endeavors in few strings, are bound to post incidentally, and are more fruitful at gathering reactions from different clients. Additionally, it is found that standoffish conduct is exacerbated when local area input is excessively brutal. Finally, an AI based strategy is used to identify disdain discourse on online client remarks from two spaces and a corpus of client remarks commented on for oppressive language.

III. TEXT CLASSIFICATION USING MACHINE LEARNING

The recent emergence of offensive language in user-generated online content has taken on increasing importance. The majority of current business techniques include boycotts and regular speech, however these tactics fall short when compared to more subdued, less clumsily delivered examples of contemptuous talk. In this study, we develop an AI-based method that outperforms a cutting-edge deep learning methodology for identifying hate speech in online client comments from two spaces.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

We also maintain a corpus of customer comments that have been reviewed for abusive language, the first of its kind. Finally, in order to deepen our understanding of this behavior, we use our discovery tool to look at harmful language across time and in different contexts.

A. Classification of Text Documents Using the Least Square Support Vector Machines with Singular Value Decomposition

Text order has emerged as one of the most important strategies for handling and coordinating content information as a result of the rapid proliferation of online data. Making up scientific categories for reports is one way to facilitate finding relevant archives, content sorting, and point tracking. The classifier used in this study to efficiently arrange text archives is LS-SVM. Text data is typically a high-dimensional trademark, thus using SVM to reduce the high-dimensionality is also possible. In this study, Least Square Support Vector Machines and Singular Value Decomposition are used to increase grouping accuracy and reduce the dimensionality of a large book's material.

Machine type: Personal details enlarged

Numerous stages try to investigate the mystery because of the harm that individual attacks on web communication create. However, it is still surprisingly difficult to comprehend the prevalence and impact of individual attacks on internet stages at scale. The goal of this work is to develop and define a method for analyzing individual assaults at scale by combining public support with AI. We demonstrate an evaluation method for a classifier that can estimate the total number of group workers. Applying our method to English Wikipedia, we get a corpus of over 100,000 excellent human-marked comments and 63,000,000 machine-named ones from a classifier that is roughly as good as the sum of three group laborers, as measured by the space under the the Spearman link and the ROC bend. Our method enables us to look at some of the unanswered questions surrounding the concept of online individual assaults using this corpus of machine-named scores. This reveals that the majority of individual attacks on Wikipedia are not the result of a small number of evil users, nor are they primarily the result of allowing enigmatic commitments from unregistered users.

IV. SYSTEM ANALYSIS

A sensible interaction is examination. Making a precise decision about how to handle the situation is the aim of this step. In order to create a consistent model of framework, tools like Class Diagram, Sequence Diagram, information stream outlines, and information word reference are used.

A. Analyzing The Domain

A computer programmer learns foundational information through space exploration, which helps them comprehend the problem. The word space used to describe the circumstance alludes to the broad area of commerce or innovation in which the client expects to use the goods. For this project, the coworkers' individual programming experiences with competing programming were considered as helping them understand the area.

B. Existing system

Online media objections are used to gradually disseminate a significant amount of material. Sadly, due to the existence of destructiveness on the internet, this vast amount of data is adversely influencing people's presences as well as the idea of human life in general.

Because harsh comments prevent people from expressing their thoughts and having refuting presumptions, there is a lack of robust dialogue via online media objections as a result of this cynicism. Determining and limiting the thoughtful direct over online discussion social events is therefore essential. Despite this, there have been attempts in the past to build internet security by page control and openly supporting.

C. Disadvantages

We have a Multi-Label Classification challenge to tackle in our data set because our data value can belong to 0 categories, 1, or more.

In order to address the issue of text categorization, this work uses six machine learning techniques: logistic regression, random forest, SVM classifier, naive bayes, decision tree, and KNN classification.

Here, we will apply six machine learning algorithms to our data in order to solve the text classification problem and choose the top algorithm based on our assessment metrics for classifying harmful remarks.

D. Proposed system

We must categorize the information in this project into six categories, such as danger, affront, hazardous, extremely poisonous, indecent, or personality contempt. We can also place one information esteem into each category.In this assignment, we must classify the information into six categories, such as danger, affront, harmful, extremely poisonous, indecent, or personality scorn. We can classify material according to its value into none of these categories, or at least more than one. Our initial assignment will be to determine whether our arrangement is multiclass or multi-mark in nature before we start any preparation on our information. A particular sketch of a nursery may contain a tree, landmark, strolling path, or a combination of these, and in this way the sketch can have a spot with nothing, at least one than one classes. This is known as multi-mark grouping.

E. Advantages

To protect teenagers, a parser and lexical feature has been combined. This allowed for the detection of harmful language in YouTube comments.

The data must then be cleaned in order to extract key features, which is the following step in our technique.

Following a thorough cleaning process, we will do an exploratory visualization to identify key features.

V. MODEL BUILDING

Describe the overall framework programming and association in this paper. Include a list of programming modules (this could include functions, subroutines, or classes), codings, and programming PC enabled programming devices (with a clear depiction of everything's capacity). Use object-arranged graphs or organised association charts to display the various division levels from highest to lowest. The graphs' highlights should all include names and references. Include a narrative that advances and improves the reader's understanding of the practical breakdown. Use subsections to address each module if appropriate.

Architecture of internal communications

Describe the system's overall communications in this area, including LANs, buses, etc. Include any implemented communication architectures, such as X.25, Token Ring, etc. Use subsections, as necessary, to discuss each architecture in use.

A. Database and File Design

During setup, collaborate with the DBA (database administrator). The section should reveal the final layout of all data base administration framework (DBMS) records and non-DBMS documents associated with the framework as a work in progress. For the particular project, further information may be added.

Give a comprehensive information word reference that includes the name, kind, length, source, approval guidelines, support (CRUD capacity—make, read, update, erase), information storage, yields, assumed names, and portrayal of each information component. is applicable as an index.

B. Files For Database Management Systems

An accurate representation of the DBMS diagrams, substructures, records, sets, tables, accumulating page sizes, etc.

Access methods (such as arranged pointer exhibit, filed, by means of set, successive, arbitrary access, and so on)

Measure the DBMS record size or the amount of information contained in the document, as well as the information pages, taking into account any overhead caused by access methods and available space.

Calculate the number of exchanges if the data set is an online exchange based framework on-Database Management System Files by determining the meaning of the update recurrence of the information base tables, views, documents, regions, records, sets, and information pages. Include a narrative description of each file's usage in this section, along with information about its usage (such as whether it is used for input, output, or both), whether it is a temporary file, which modules read and write to it, and its file structure (refer to the data dictionary). The information about the file structure should:

Recognise the record structures, record keys or indexes, and the references to the records' data elements.

Record length (fixed or maximum variable length) and blocking variables should be defined.

Define the file access mechanism, such as random access, virtual sequential access, index sequential access, etc.

Calculate the size of the file or the amount of data it contains, taking file access method overhead into account.

Calculate the size of the file or the amount of data it contains, taking file access method overhead into account.

Provide the projected number of transactions per unit of time, along with the statistical mean, mode, and distribution of those transactions, if the file is a component of an online transaction-based system, and define the update frequency of the file.

C. Environmental Interface

Regardless of whether other frameworks are supervised by the State or another office, outer frameworks are any frameworks that are beyond the scope of the framework being worked on. Show the electronic interfaces between this framework and each and every other framework as well as any subsystems in this section, putting emphasis on the viewpoint of the framework that is being developed.

Architecture for Interfaces

Show the interfaces between the framework being used in this segment.Describe the interface(s) between the framework being developed and other frameworks in this section, such as group motions, inquiries, and so on. Include any interface architectures that are currently being used, such as wide region organizations, passageways, and so on. Provide a diagram that shows the relationship between this framework and each of the other frameworks, showing how it relates to the setting outlines. Use subsections to address each active interface, if that is appropriate.

Details of the interface design

There must be rules governing the interface for any framework that provides data exchange with the framework under development. This section should include enough specific information about the requirements of the interface to correctly arrange, convey, and also obtain information via the interface.

Keep in mind the following information for the specific strategy for each interface (where appropriate):

The requirements for information design are: Instruments and methods for the reformat interaction should be described if it is necessary to reformat information before it is conveyed or after it has been acquired.

Details for hand-shaking protocols between the two frameworks, including the nature and format of the information to be recalled for the hand-shake messages, the context for exchanging these messages, and the actions to be done when errors are identified

Format(s) for error reports exchanged across frameworks should take into account the aura of error reports; for example, they could be kept in a record or printed out and sent to the administrator.

A visual representation of the availability between frameworks that displays the direction of the information stream depictions of questions and responses

The information can be copied from an official Interface Control Document (ICD) for a specific interface, or the ICD can be mentioned in this section. regulating system integrity

Touchy frameworks use information whose misuse, exploitation, modification, or unauthorised access may affect the direction of State programmes or the level of protection to which individuals are legally entitled.

Developers of sensitive State systems must create requirements for at least levels of controls the following Internal security to restrict access to only those entrance types required by clients and to basic information items

Examine methods for ensuring that operational and board reports satisfy the requirements for control, detailed, and maintenance periods. strong application review trails to examine recovery admission to given fundamental information Tables that must be used or specified in order for information fields to be approved Ability to identify all review data by client ID, network terminal recognisable proof, date, time, and information got to or altered. Confirmation measures for increases, cancellations, or changes of basic data.

VI. MODULES

The best progression of the "Web" of the 21st century may have been made possible by the exceptional advancement of software engineering and innovation, which allows one person to impart to another globally with the help of a computer. Exponential growth: The amazing advancement of software engineering and innovation has given rise to the "Web" of the twenty-first century, which allows one person to transmit to another globally with the use of a basic PDA and the internet. People used to communicate with one another in the early days of the internet, namely through Email, which was flooded with spam. Ordering the messages back then as positive or negative, such as spam or not-spam, was a significant task. Since the advent of online media outlets, correspondence and the flow of information through the internet have undergone significant change.With the development of online media, it has become increasingly important to categorize the content into positive and negative categories in order to prevent social harm and regulate people's reserved behavior.

A. Machine learning

On the Kaggle.com data set, various machine learning methods will be employed to categorize harmful comments. This work uses machine learning methods, such as logistic regression and naive bayes, to address the text categorization issue.

So, using the provided data set, we will apply each machine learning technique, calculate and compare each algorithm's accuracy, log loss, and hamming loss. Since binary data is what computers use, data in the actual world can take on many different forms, such as text or graphics. As a result, in order for the data from the actual world to be properly processed by the computer, it must be converted into binary form. In this research, we will classify online comments using this converted data and machine learning approaches.

By applying the data to a function that will assign a value to each data value of the data set, text classification may be applied to a given data set and set of labels with ease.

B. Convolution neural network

Different AI calculations will be used to order the harmful comments on the Kaggle.com data set. In order to address the problem of text characterization, this study includes six AI techniques, such as strategic relapse, arbitrary timberland, SVM classifier, credulous bayes, decision tree, and KNN grouping. In order to analyse the accuracy, log misfortune, and hamming misfortune of all six AI calculations, we will apply them to the informational index and figure provided. In the real world, we have information in many structures, such as pictures or text, as computers work on parallel information. As a result, we must convert the data from the current reality into a parallel structure for proper PC preparation. In this research, we will make use of this updated data and use machine learning techniques to classify online comments. By applying the information to a capacity that will designate a worth to every piece of information in the informational collection, text order may be conveniently applied to a particular informational collection and set of names.

C. Finishing Up Evaluation Metrics: Evaluation

The nature of AI calculations is verified through assessment measurements. Therefore, in order to determine and evaluate all of the strategies, we need to choose the appropriate assessment metrics for our informational index before applying any AI computations to the handled data. There are two important categories of measures for multi-name arrangements. Metrics Based on Models: Here, we'll calculate the incentive for each informational value before averaging the results across the informational index. Accuracy, model hamming loss, and other factors Name-Based Metric: In this case, we will calculate the incentive for each name in our order and then average out all of the attributes without accounting for any correlation between marks. Model typical accuracy, a single error, and so forth.

VII. FUTURE ENHANCEMENTS

For improved results, subsequent research can use different AI models to calculate exactness, hamming misfortune, and log misfortune. We may also look into some complex learning algorithms as the GRU, multi-facet perceptron, and LSTM (long transient memory repetitive neural organisation). This allows us to research a wide range of techniques that will help us improve the outcome.

VIII. CONCLUSION

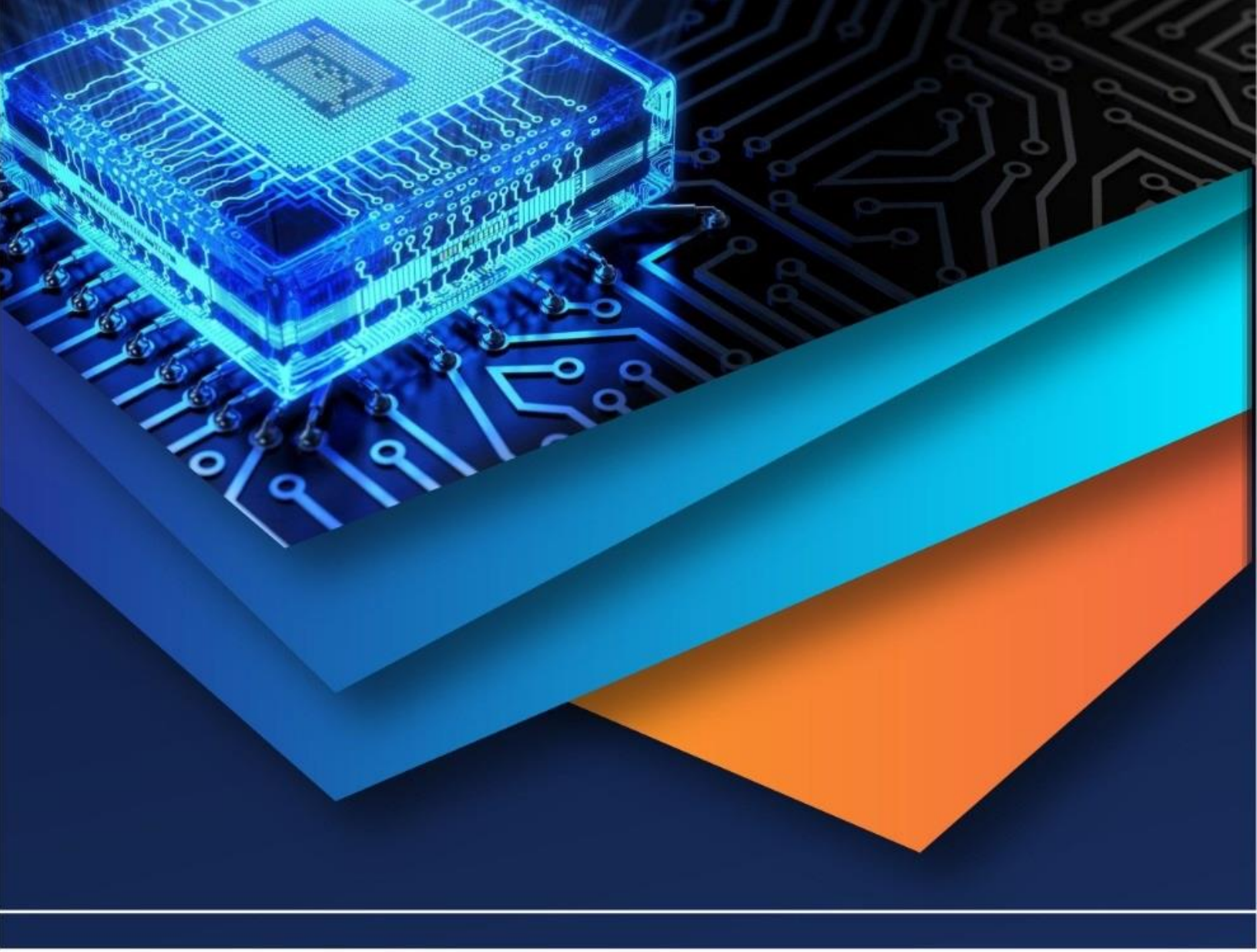
In this study, we compare the hamming loss, accuracy, and log loss of three machine learning techniques: logistic regression, Naive Bayes, and multi-label classifier. After thorough analysis, we can now state that logistic regression performs best in terms of hamming loss because our hamming loss is lowest in that case, while logistic regression performs best in terms of accuracy because accuracy in that model is best compared to others, and random forest performs best in terms of log loss because it has the lowest possible log loss in that model. So, hamming loss and accuracy will be combined to determine the final model we choose. Since we obtained the highest accuracy (96.46%) and the lowest hamming loss (2.43%) in the instance of the logistic regression. Since the logistic regression model performs the best for our data, we will choose it as our final machine learning method.

IX. ACKNOWLEDGMENTS

I would like to express my special thanks and gratitude to my college(G Narayanamma institute of technology and science) and my department, for encouraging me to research on this project, which also helped me in gaining a lot of practical knowledge and I came to know about so many new things theoretically, and also helped in mastering certain useful topics, I am really thankful to them. Secondly, I would also like to thank my parents and friends who helped me a lot in finalizing this paper within the limited time frame.

REFERENCES

[1] K Ramya Sri and Ch Ramesh “Credit Risk Valuation Using Machine Learning Algorithm” Springer Learning and Analytics in Intelligent Systems series, vol. 4, chapter 74, June 2019 https://link.springer.com/chapter/10.1007%2F978-3-030-24318-0_74
[2] Ch Ramesh and K Ramya Sri, "Evaluation of Machine Learning Models for Credit Scoring[J]", Test Engineering and management, vol 82, Page number 2798-2805. ISSN 0193-4120, https://zenodo.org/record/7920846
[3] M. A. Walker, P. Anand, J. E. F. Tree, R. Abbott, and J. Lord, "A corpus for research on pondering and discussion," Proc. eighth Int. Conf. Lang. Resour. Eval. Lr. 2012, pp. 812 – 817, 2012.
[4] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "Withdrawn conduct in online conversation networks," Proc. ninth Int. Conf. Web Soc. Media, ICWSM 2015, pp. 61–70, 2015.
[5] B. Mathew et al., "Thou shalt not disdain: Countering on the web disdain discourse," Proc. thirteenth Int. Conf. Web Soc. Media, ICWSM 2019, no. August, pp. 369–380, 2019.
[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Harmful language discovery in online client content," 25th Int. Internet Conf. WWW 2016, pp. 145 – 153, 2016, doi: 10.1145/2872427.2883062.
[7] E. K. Ikonomakis, S. Kotsiantis, and V. Tampakas, "Text Classification Using Machine Learning Techniques," no. August 2005.
[8] M. R. Murty, J. V... Murthy, and P. Reddy P.V.G.D, "Text Document Classification basedon Least Square Support Vector Machines with Singular Value Decomposition," Int. J. Comput. Appl., vol. 27, no. 7, pp. 21–26, 2011, doi: 10.5120/3312-4540.
[9] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal assaults seen at scale," 26th Int. Internet Conf. WWW 2017, pp. 1391–1399, 2017, doi: 10.1145/3038912.3052591.
[10] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, "Misleading Google's Perspective API Built for Detecting Toxic Comments," 2017, [Online]. Accessible: http://arxiv.org/abs/1702.08138.
[11] Y. Kim, "Convolutional neural organizations for sentence arrangement," EMNLP 2014 - 2014 Conf. Empir. Strategies Nat. Lang. Interaction. Proc. Conf., pp. 1746–1751, 2014, doi: 10.3115/v1/d14 - 1181.
[12] R. Johnson and T. Zhang, "Successful utilization of word request for text order with convolutional neural organizations," NAACL HLT 2015
[13] Y. Chen and S. Zhu, "Recognizing Offensive Language in Social Media to Protect Adolescents," [Online].Available: http://www.cse.psu.edu/~sxz16/papers/SocialCom2012.pdf.
[14] A. L. Sulke and A. S. Varude, "Order of Online Pernicious Comments utilizing Machine Learning," no. October 2019
[15] I.Ravi Prakash Reddy, “Location Sharing System with Enhanced Privacy in Mobile Online Social Network”, Journal of Emerging Technologies and Innovative Research, ISSN-2349-5162, Vol-6, Issue-6, June-2019
[16] N. Chakrabarty, "A Machine Learning Approach to Comment Toxicity Classification," Adv. Intell. Syst. Comput., vol. 999, pp. 183 – 193, 202
[17] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis and Vassilis P. Plagianakos:” Convolutional Neural Networks for Toxic Comment Classification”
[18] Kevin Khieu and Neha Narwal:” Detecting and Classifying Toxic Comments”, https://web.stanford.edu/class/cs224n/report s/6837517.pdf
[19] Vaddi, S., Mohanty, H. (2014). Web Service Composition Using Service Maps. In: Murty, M.N., He, X., Chillarige, R.R., Weng, P. (eds) Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2014. Lecture Notes in Computer Science(), vol 8875. Springer, Cham. https://doi.org/10.1007/978-3-319-13365-2_18
[20] Theodora Chu, Kylie Jue and Max Wang:” Comment Abuse Classification with Deep Learning”, https://web.stanford.edu/class/cs224n/reports/2762092.pdf
[21] Sesha Bhargavi, V., Spandana, T. (2017). Recommendation Based P2P File Sharing on Disconnected MANET. In: Deiva Sundari, P., Dash, S., Das, S., Panigrahi, B. (eds) Proceedings of 2nd International Conference on Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 467. Springer, Singapore. https://doi.org/10.1007/978-981-10-1645-5_18
[22] Velagaleti, Sesha Bhargavi, M. Seetha, and S. Viswanadha Raju. "A Simulation and Analysis Of Secured AODV Protocol in Mobile Ad Hoc Networks." IJRET International Journal of Research in Engineering and Technology (2013).
[23] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data
[24] Coversation AI Team, https://conversationai.github.io/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY