CYBER BULLYING DETECTION: AN ENSEMBLE BASED MACHINE LEARNING APPROACH

S. APARNA, R. PRIYANKA, CH. NAVEEN KUMAR, P. HEMANTH KIRAN KUMAR

Under The Esteemed Guidance Of Mrs. J. Lakshmi

Department of Computer Science and Engineering

BACHELOR OF TECHNOLOGY

TIRUMALA ENGINEERING COLLEGE Jonnalagadda, Narasaraopet, GUNTUR (Dt.), A.P.

ABSTRACT

Research on cyberbullying detection.To identify control and reduce thebullying contents spread over social media sites. Though data collection andfeature engineering process has been elaborated, yet most of the emphasis ison feature selection algorithms. Thus there is an extensive need to identify, control and reduce the bullying contents spread over social media sites, which hasmotivated us to conduct this research to automate the detection process of offensive language or cyberbullying.

In our work, Logistic Regression and Bagging ensemble model classifier have performed individually best in detecting cyberbullying which has been outperformed by our proposed SLE and DLE voting classifiers. Our proposed SLE and DLE models yield the best performance of 96% when TF-IDF (Unigram) feature extraction is applied with K-Fold cross-validation.

introduction

Machine or deep learning algorithms help researchers understand big data.Abundant information on humans and their societies can be obtained in this bigdata era, but this acquisition was previously impossible .One of the main sources of human-related data is social media (SM). By applying machine learning algorithms to SM data, we can exploit historical data to predict the future of a wide range of applications. Machine learning algorithms provide an opportunityto effectively predict and detect negative forms of human behavior, such as cyberbullying.

The big data analysis can uncover hidden knowledge through deep learning from raw data . Big data analytics has improved several applications, andforecasting the future has even become possible through the combination of big data and machine learning algorithms. An insightful analysis of data on human behavior and interaction to detect and restrain aggressive behavior involves multifaceted angles and aspects and the merging of theorems and techniques frommultidisciplinary and interdisciplinary.

The accessibility of large-scale data produces new research questions, novel computational methods, interdisciplinary approaches, and outstanding opportunities to discover several vital inquiries quantitatively. However, using traditional methods (statistical methods) in this context is challenging in terms ofscale and accuracy. These methods are commonly based on organized data on human behavior and small-scale human networks (traditional social networks).Applying these methods to large online social networks (OSNs) in terms of scaleand extent causes several issues. On the one hand, the explosive growth of OSNsenhances and disseminates aggressive forms of behavior by providing platformsand networks to commit and propagate such behavior.

On the other hand, OSNs offer important data for exploring human behavior and interaction at a large scale, and these data can be used by researchers to develop effective methods of detecting and restraining misbehavior and/or aggressive behavior. OSNs provide criminals with tools to perform aggressive actions and networks to commit misconduct. Therefore, methods that address bothaspects (content and network) should be optimized to detect and restrain aggressive behavior in complex systems

2. LITERATURE SURVEY

2.1 .TITLE: Predicting human behavior: The next frontirers.

AUTHOUR: V. Subrahmanian and S. Kumar.

ABSTRACT: Machine learning has provided researchers with new tools for understanding human behavior. In this article, we briefly describe some successesin predicting behaviors and describe the challenges over the next few years.

2.2 .TITLE: Homophily in the digital world: A LiveJournal case study. **AUTHOUR**: H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas. **ABSTRACT**: Are two users more likely to be friends if they share common interests? Are two users more likely to share common interests if they're friends?The authors study the phenomenon of homophily in the digital world byanswering these central questions. Unlike the physical world, the digital worlddoesn't impose any geographic or organizational constraints on friendships. So, although online friends might share common interests, a priori there's no reason to believe that two users with common interests are more likely to be friends. Using data from LiveJournal, the authors show that the answer to both questions is yes.

2.3 TITLE: Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitternetwork.
AUTHOUR: M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana.

ABSTACT: The popularity of online social networks has created massive social communication among their users and this leads to a huge amount of user-generated communication data. In recent years, Cyberbullying has grown into amajor problem with the growth of online communication and social media. Cyberbullying has been recognized recently as a serious national health issue among online social network users and developing an efficient detection modelholds tremendous practical significance. In this paper, we have proposed set of unique features derived from Twitter; network, activity, user, and tweet content,based on these feature, we developed a supervised machine learning solution for detecting cyberbullying in the Twitter.

An evaluation demonstrates that our developed detection model based onour proposed features, achieved results with an area under the receiver-operatingcharacteristic curve of 0.943 and an f-measure of 0.936. These results indicate that the proposed model based on these features provides a feasible solution todetecting Cyberbullying in online communication environments. Finally, we compare result obtained using our proposed features with the result obtained from two baseline features. The comparison outcomes show the significance of the proposed features.

. This paper presents a new methodologyfor automating social media data, and
for OSN users, which can improve their securityand privacy when using these platforms. Furthermore, we suggest future researchdirections.

Features of Python Programming

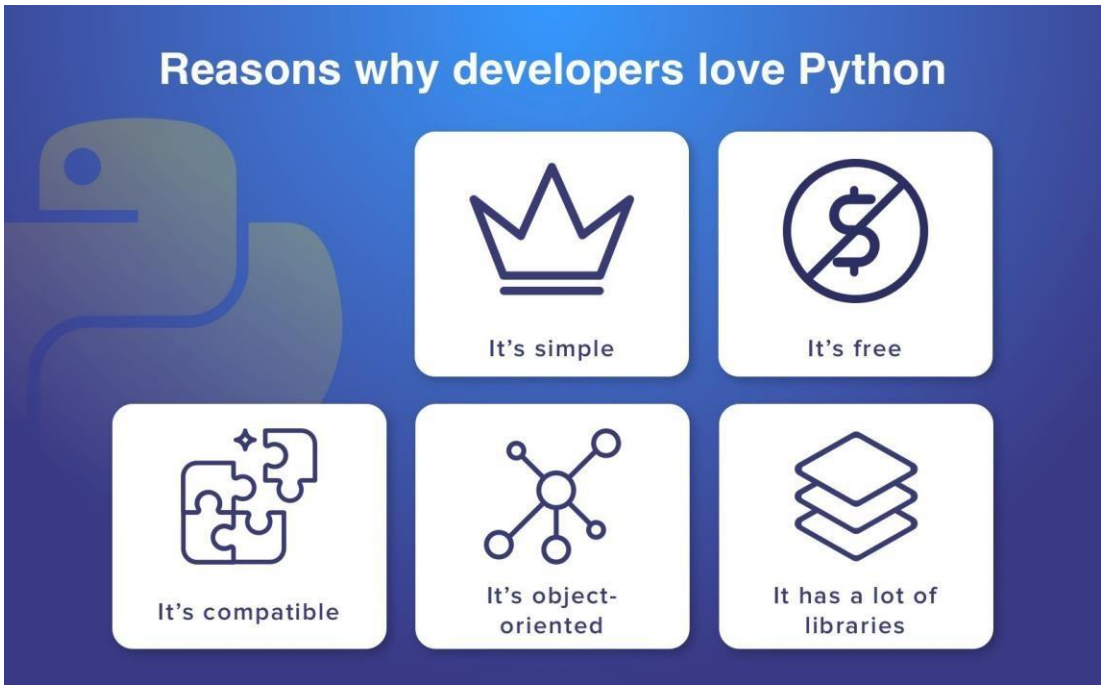


Figure 3.1. The Features of Python Programming

1. A simple language which is easier to learn

- Python has a very simple and elegant syntax.
- It's much easier to read and write Python programs compared toother languages like: C++, Java, C#.
- Python makes programming fun and allows you to focus on thesolution rather than syntax.
- If you are a newbie, it's a great choice to start your journey withPython.

2. Free and open-source

- You can freely use and distribute Python, even for commercial use.
- Not only you can use and distribute software's written in it, youcan even make changes to the Python's source code.
- Python has a large community constantly improving it in eachiteration.

PYTHON PROGRAM TO ADD TWO NUMBERS

# This program adds two numbersnum1 = 1.5
num2 = 6.3

Add two numbers

sum = float(num1) + float(num2)# Display the sum
print('The sum of {0} and {1} is {2}'.format(num1, num2, sum))

ADD TWO NUMBERS PROVIDED BY THE USER

```
# Store input numbers
num1 = input('Enter first number:
') num2 = input('Enter second
number: ')


# Add two numbers
sum = float(num1) + float(num2)


# Display the sum
```

We use the built-in function input() to take the input.

input() returns a string, so we convert it into number using the float() function.

Python Quick start

Python is an interpreted programming language, this means that as a developeryou write Python (.py) files in a text editor and then put those files into the pythoninterpreter to be executed.

The way to run a python file is like this on the command line:C:\Users*Your Name*>python helloworld.py
Where "helloworld.py" is the name of your python file.

Let's write our first Python file, called helloworld.py, which can be done in anytext editor.

helloworld.py print("Hello, World!")
Simple as that. Save your file. Open your command line, navigate to the directorywhere you saved your file, and run:

C:\Users*Your Name*>python helloworld.pyThe output should read:
Hello, World!

Congratulations, you have written and executed your first Python program.

OTHER TESTING METHODOLOGIES

User Acceptance Testing

User Acceptance of a system is the key factor for the success of any system. Thesystem under consideration is tested for user acceptance by constantly keeping in touch with theprospective system users at the time of developing and making changes wherever required. Thesystem developed provides a friendly user interface that can easily be understood even by a person who isnew to the system.

OUTPUT TESTING

After performing the validation testing, the next step is output testing of theproposed system, since no system could be useful if it does not produce the required outputin the specified format. Asking the users about the format required by them tests theoutputs generated or displayed by the system under consideration. Hence the output format isconsidered in 2 ways – one is on screen and another in printed format.Validation Checking
Validation checks are performed on the following fields.

Text Field:

The text field can contain only the number of characters lesser than or equal to itssize.

The text fields are alphanumeric in some tables and alphabetic in other tables.
Incorrect entry

always flashes and error message.

Numeric Field:

The numeric field can contain only numbers from 0 to 9. An entry of anycharacter flashes an error messages. The individual modules are checked for accuracy andwhat it has to perform. Each module is subjected to test run along with sample data. Theindividually

tested modules are integrated into a single system. Testing involves executing thereal data information is used in the program the existence of any program defect is inferredfrom the output. The testing should be planned so that all the requirements are individuallytested.

A successful test is one that gives out the defects for the inappropriate data andproduces and output revealing the errors in the system.

Preparation of Test Data

Taking various kinds of test data does the above testing. Preparation of test dataplays a vital role in the system testing. After preparing the test data the system understudy is tested using that test data. While testing the system by using test data errors are againuncovered and

corrected by using above testing steps and corrections are also noted for futureuse.

Using Live Test Data:

Live test data are those that are actually extracted from organization files. After a system is partially constructed, programmers or analysts often ask users to key ina set of data from their normal activities. Then, the systems person uses this data as a way topartially test the system. In other instances, programmers or analysts extract a set of live datafrom the files and have them entered themselves.

It is difficult to obtain live data in sufficient amounts to conduct extensive testing.And, although it is realistic data that will show how the system will perform for thetypical processing requirement, assuming that the live data entered are in fact typical, such datagenerally will not test all combinations or formats that can enter the system. This bias toward typicalvalues then does not provide a true systems test and in fact ignores the cases most likely tocause system satisfy the needs to the largest possible extent. With development in technology, itmay be possible to add many more features based on the requirements in future. The coding anddesigning is simple and easy to understand which will make maintenance easier.

TESTING STRATEGY :

A strategy for system testing integrates system test cases and design techniquesinto a well planned series of steps that results in the successful construction of software. The testingstrategy must co-operate test planning, test case design, test execution, and the resultant datacollection and evaluation

.A strategy for software testing must accommodate low-level tests that arenecessary to verify that a small source code segment has been correctly implemented as well as high leveltests that validate major system functions against user requirements.

Software testing is a critical element of software quality assurance and representsthe ultimate review of specification design and coding. Testing represents an interesting anomaly forthe software. Thus, a series of testing are performed for the proposed system before the system is readyfor user acceptance testing.

SYSTEM TESTING:

Software once validated must be combined with other system elements (e.g.Hardware, people,

database). System testing verifies that all the elements are proper and that overallsystem function performance is achieved. It also tests to find discrepancies between the systemand its original objective, current specifications and system documentation.

UNIT TESTING:

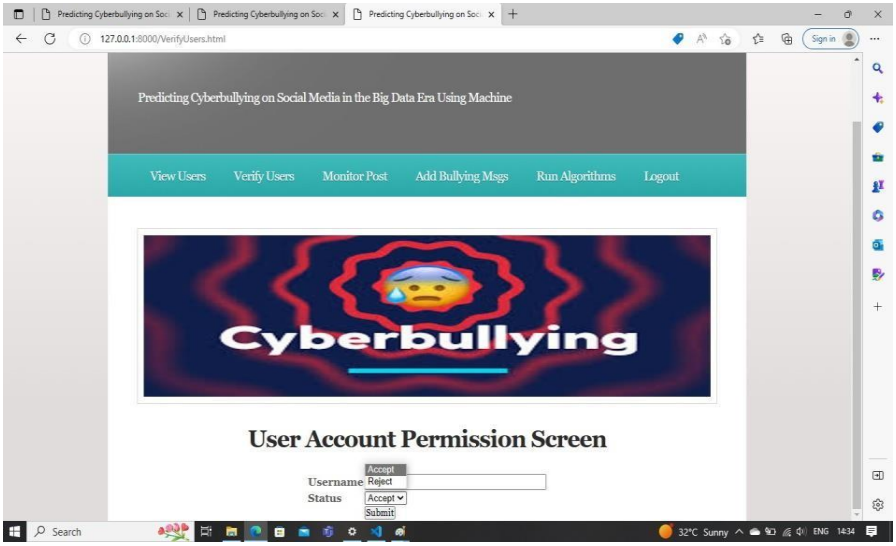
In unit testing different are modules are tested against the specifications producedduring the design for the modules. Unit testing is essential for verification of the code produced duringthe coding phase, and hence the goals to test the internal logic of the modules. Using the detailed designdescription as a guide, important Conrail paths are tested to uncover errors within the boundary ofthe modules. This testing is carried out during the programming stage itself. In this type of testingstep, each module was found to be working satisfactorily as regards to the expected output from themodule.

In Due Course, latest technology advancements will be taken into consideration.
As part of

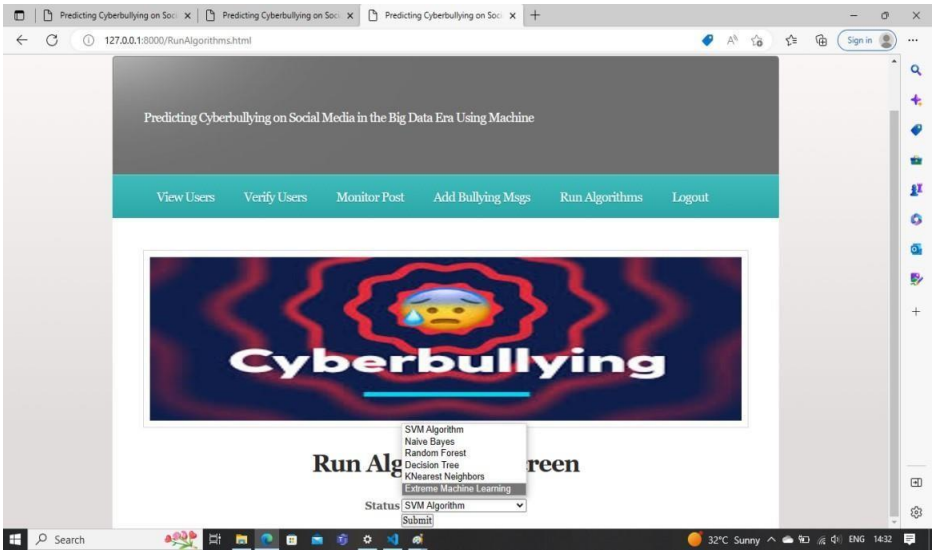
technical build-up many components of the networking system will be generic innature so that future projects can either use or interact with this. The future holds a lot to offer to thedevelopment and refinement of this project.

Admin Login

User Accepted by the Admin



Algorithms



CHAPTER-8 CONCLUSION

CONCLUSION

This study reviewed existing literature to detect aggressive behavior on SM websites by using machine learning approaches. We specifically reviewed four aspects of detecting cyberbullying messages by using machine learning approaches, namely, data collection, feature engineering, construction of cyberbullying detection model, and evaluation of constructed cyberbullying detection models. Several types of discriminative features that were used to detectcyberbullying in online social networking sites were also summarized. In addition, the most effective supervised machine learning classifiers for classifying cyberbullying messages in online social networking sites were identified. One of the main contributions of current paper is the definition of evaluation metrics to successfully identify the significant parameter so the various machine learning algorithms can be evaluated against each other. Most importantly we summarized and identified the important factors for detecting cyberbullying through machine learning techniques specially supervised learning. For this purpose, we have used accuracy, precision recall and f-measure which gives usthe area under the curve function for modeling the behaviors in cyberbullying.Finally, the main issues and open research challenges were described and discussed.

REFERENCES

1. V. Subrahmanian and S. Kumar, ``Predicting human behavior: The nextfrontiers," Science, vol. 355, no. 6324, p. 489, in 2017.
2. H. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, ``Homophily in thedigital world: A LiveJournal case study," IEEE Internet Comput., vol. 14,no. 2, pp. 15_23, in Mar./Apr. 2010.
3. M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, ``Cybercrime detectionin online communications: The experimental case of cyberbullying detection in the Twitter network," Comput. Hum. Behav., vol. 63, pp. 433_443, in Oct. 2016.
4. L. Phillips, C. Dowling, K. Shaffer, N. Hodas, and S. Volkova, ``Using social media to predict the future: A systematic lite rature review," 2017,arXiv:1706.06134. [Online]. Available: https://arxiv.org/abs/1706.06134
5. H. Quan, J. Wu, and Y. Shi, ``Online social networks & social network services: A technical survey," in Pervasive Communication Handbook.Boca Raton, FL, USA: CRC Press, in 2011.
6. J. K. Peterson and J. Densley, ``Is social media a gang? Toward a selection,facilitation, or enhancement explanation of cyber violence," Aggression Violent Behav., in 2016.
7. BBC. (2012). Huge Rise in Social Media. [Online]. Available:http://www.bbc.com/news/uk-20851797
8. P. A.Watters and N. Phair, ``Detecting illicit drugs on social media using Automated social media intelligence analysis (ASMIA)," in CyberspaceSafety and Security. Berlin, Germany: Springer, 2012, pp. 66_76.
9. M. Fire, R. Goldschmidt, and Y. Elovici, ``Online social networks: Threats and solutions," IEEE Commun. Surveys Tuts., vol. 16, no. 4,pp. 2019_2036, 4th Quart., in 2014.
10. N. M. Shekokar and K. B. Kansara, ``Security against sybil attack in socialnetwork," in Proc. Int. Conf. Inf. Commun. Embedded Syst. (ICICES), in 2016.
11. J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, ``Detecting and tracking political abuse in social media.