DETECTION OF CYBERBULLY USING MACHINE LEARNING APPORACH

divya[1], usr praneeth [2],

[1]**Assistant Professor MSC dept,** Dantuluri Narayana Raju College**, Bhimavaram, Andharapradesh**
[2]**PG Student of MSC, Dantuluri** Narayana Raju College**, Bhimavaram, Andharapradesh**

Abstract— *Cyberbullying is a type of harassment that is carried out through the use of technology. While this has long been an issue, the impact on young people has just recently become obvious. Threats thrive on social networking platforms, which are also targets for attacks by young people who use them. We can build standards for automated cyberbullying content identification by using algorithms to identify bully and victim language trends. The data for our study came from the website Formspring.me, which is a question-and-answer platform containing a lot of harassing content. Amazon's Mechanical Turk cloud service was used to name the data. We employed python's AI strategies in combination with python's AI methodologies.*

Index Terms— unigram, n-gram, bigram, trigram, Tf-Idf, Bully, Non-Bully

I Introduction

Cyberbullying is defined as the use of the Internet, mobile phones, or other technologies to communicate or upload content or images with the intent to injure or humiliate another person, according to the National Crime Prevention Council. StopBullying.gov defines cyberbullying as "bugging that occurs through electronic advancement (joins) equipment and apparatus, such as mobile phones, PCs, and tablets, as well as correspondence stages such as online life goals, messages, talking, and places." It's possible that the survivor and a few other people will see it again. As a result, the victim may be tormented multiple times. Because of the torment that takes place on the internet, it will, in general, become less useful.

Name-calling

Rumor-mongering is the act of distributing false or malicious information about someone in order to create a rumour.

Uploading photos of someone that are lewd or embarrassing Terrorism and threats

Posting private, personal, or humiliating data, images, or recordings in a public place is known as outing.

Trolling is the act of persuading people to react enthusiastically in an online network by utilising affronts or provocative language.

Cyberstalking is the practise of following or harassing someone who uses the Internet or other electronic devices. It includes things like watching, terrorising, and the threat of harm, as well as hostile words that can range from threats of bodily harm to the targeted casualty, and so on..

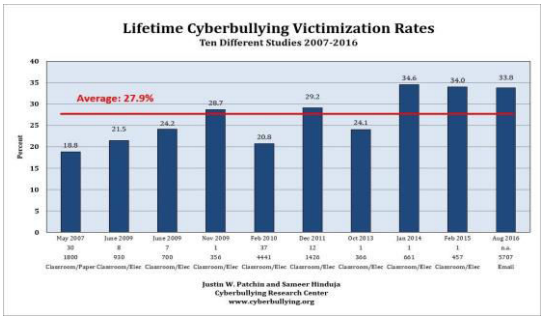

Fig 1 shows rates of cyber bullying victimization have varied between 2007 and2016.

2 Literature survey

[1 Chen et al. investigate the detection of offensive language in social media using the lexical syntactic feature (LSF) approach, which successfully detects offensive material and users who send offensive messages in social media.

. [2] The researchers at Dinakar et al. are interested in detecting textual cyberbullying in YouTube comments. They gathered videos that dealt with difficult issues such as race and culture, sexuality, and intelligence. It was demonstrated that binary classifiers beat multi-class classifiers by manually tagging 4,500 YouTube comments and using binary

and multi-class classifiers. A language-based method was used by Reynolds et al. to detect cyberbullying. As characteristics, the quantity, density, and value of offensive words were employed. By tracking the percentage of curse and insult in posts, their research has effectively established Type Spring Messages that contain cyberbullying.

[3 Yin et al. used three types of features to train a model for detecting harassing messages in chat rooms and discussion forums: content features (word occurrence weighted by term-frequency inverse-document-frequency, or TF-IDF), relational features (offensive terms and pronouns), and contextual features (the resemblance of a user to a neighbour). By automatically monitoring and reporting online rumours and threats to parent-teacher association (PTA) school representatives, Ptaszynski et al. offer an organised approach to internet surveillance.

[4 Initially, a dictionary of cyberbullying-related swear words was painstakingly constructed. Dani and colleagues utilise sentiment analysis to detect cyberbullying via social media. Nandini et al. exploit language features to identify cyberbullying episodes in the social network using fuzzy logic and genetic algorithms. Huang et al. study if analysing social network features might improve cyberbullying detection accuracy.

3 INPUT DATA

The relationship between the user and the information system is defined by the input design. It necessitates the development of data preparation specifications and procedures, and these measures are required to position transaction data in a usable form for processing by requiring the machine to read data from a written or printed document, or by requiring people to lock data directly into the database. The input design focuses on controlling the quantity of data required, decreasing errors, preventing delays, avoiding extra steps, and making the process simple. The input has been developed to provide security and convenience of use while maintaining privacy. The following factors were taken into account by Output Development:

considered the following:

What data should be given asinput?

- How the data should be arranged orcoded?
- The dialog to guide the operating personnel in providinginput.
- Methods for preparing input validations and steps to follow when erroroccur.

4. Implementation Study

We gathered data from the Ask.fm social networking site (http:/ask.fm), received donations, and set up simulated tests with volunteer adolescents to create a corpus. There have been 91,370 Dutch posts registered. Ask.fm The Ask.fm social networking platform, where users may establish profiles and ask and answer questions, provided a substantial portion of our corpus. It is possible to do so anonymously.

Typically, Ask.fm data is made up of question-answer pairs that are published on the user's profile. The information was gathered by utilising the GNW Get software (https://www.gnu.org/software/wget) to search a number of seed sites. After filtering out non-Dutch content, this yielded 85,462 postings.

We undertook two efforts to supplement the data set because cyberbullying posts were underrepresented in the corpus.**4 Proposed Approach**

As can be observed, the proposed strategy consists of three essential steps: Preprocessing, feature extraction, and classification are all steps in the classification process. We sanitise the data in the preprocessing step by removing noise and superfluous text. The preprocessing procedure is as follows:

- Tokenization: In this section, we take the text as sentences or paragraphs and output it as a list of separated words.
- Lowering text: This lowers all of the letters in the list of words that escaped tokenization. 'THIS IS AWESOME', for example, will be 'this is awesome.'
- Stop words and encoding cleaning: This is an

important step in the preprocessing process where we remove stop words and encoding characters like n and t from the text.

The features extraction stage is the second step in the suggested Model. The textual input is translated into a format that can be fed into machine learning algorithms in this step. First, we use TFIDF to extract the features from the input data and store them in a features list. TFIDF's main notion is that it works on the text to determine the weights of words in relation to the document resentence.

We apply a sentiment analysis technique in addition to TFIDF to extract the polarity of the phrases and add it as a feature to the features list including TFIDF features. The polarity of the sentences indicates whether or not the sentence is positive or negative.

For this, we use the Text Blob library, which is a pre-trained model on movie reviews, to extract the polarity. The proposed approach leverages N- Gram to consider alternative combinations of words during model evaluation, in addition to feature extraction using TFIDF and sentiment polarity extraction. We used 2- Gram, 3-Gram, and 4-Gram in particular.

The classification step is the final step in the proposed approach, in which the extracted features are fed into a classification algorithm to train, test, and use the classifier in the prediction phase. SVM (Support Vector Machine) and navie bayes were utilised as classifiers.

To begin, the annotators were asked to establish that a post is a portion of a digital tormenting case at the post level. This was accomplished by assigning a contamination score to the post on a three-point scale, with 0 indicating that the post contains no signs of cyberbullying, 1 indicating that the post contains minor signs of cyberbullying, and 2 indicating that the post contains significant signs of cyberbullying.

When a post was determined to be a part of a cyberbullying foundation (e.g., a hurtfulness score of 1 or 2), the annotators advised the creator's position in the cyberbullying case. Aside from the survivor and harasser, we recognise two types of observers in our explanation framework:

1) bystander protectors who support the individual in question and prevent the harasser from continuing his demonstrations;

2) bystanders who do not start or participate in the harasser's demonstrations.

Second, despite the fact that the post was not deemed damaging, the annotators were accused of describing fine-grained content classifications appropriate to cyberbullying at the level of the sub-sentence.

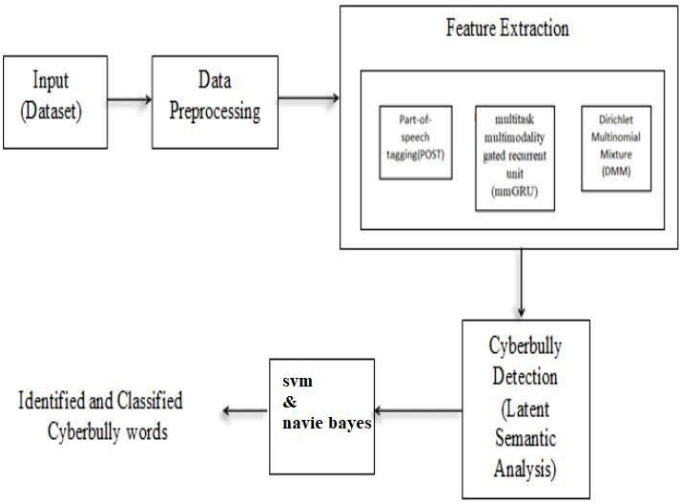

Fig 1: - proposed model

5 Algorithms Used

Algorithm

These algorithms use special terminology called F-measure, Precision, Recall to obtain the good accuracy results.

The terms can be used in this chapter are described as follows:

1) Uni gram: It represents only “one” word.
2) Bi-gram: It is a sequence of “two” words
3) Trigram: It makes prediction for the word based on two words before that.
4) N-gram: N-gram may not have any

relation between them apart form the fact they appear next to each other.

5) F-measure: F1-score or F-score is a measure of test's accuracy and is defined as the weighted harmonic mean of the precision and recall of the test.

An effective cyberbullying prediction models is to use a text classification approach that involves the construction of machine learning classifiers from labeled text instance and is to use a lexicon-based model that involves computing orientation for a document from the semantic orientation of words or phrases in the document.

ContentQuery 1	Bitch-ass skank dickhead friendless fuckoff Fuck stick kuntassfuckclitfacedumshitfaggot tt fuckface motherfucking negro tar cunt fag hoe fuckin nasty bitch douche-fag faggot nigger trash
ContentQuery 2	ass-shit, ass-fuck, ass-face, shit, bitch, fuck, ugly, ass- bite, ass-fucker, ass-cock, ass-hole, ass-hat, ass-nigger, ass-monkey, ass-clown, ass-pirate, ass-sucker, ass-wipe, ass, ass-banger, ass-cracker, ass-hopper, ass-jabber, ass- jacket, ass-licker, ass-wad, fucking, big, fake, gay, dick, stupid, hoe, pussy, damn, hell, dumb, fat, kill
ContentQuery 3	Bitch, fucking, hoe, ugly, cunt, fag, ass-shit, pussy, Dumb ass, douche-fag, nigger, trash, ass-fuck, shit, ass-fucker, ass-cock, assnigger, ass-monkey, ass-clown, ass-pirate, ass-sucker, ass-wipe, ass, ass-banger, ass cracker, ass- hopper, ass-jabber, ass-jacker, ass-licker, ass-wad, fake, dick, damn, fat, whore, fuckin, stfu, fucked, faggot, fuck, ass-bite, ass-hole, kill, nigga, bitches, fucks, loser, dick
ContentQuery 4	All terms in the bad words dictionary.

Table 2 content-based queries

6 Results and Evolution Metrics

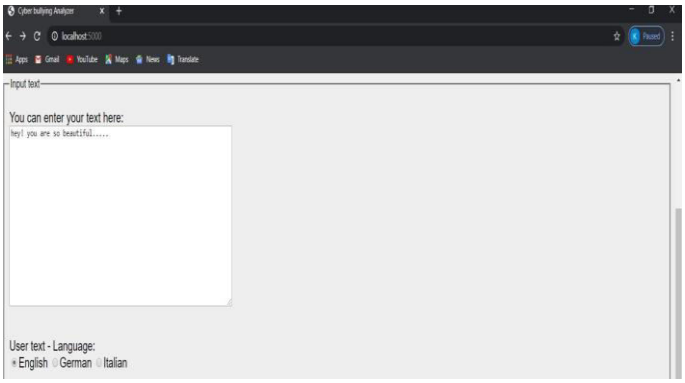

Fig 2:- **- input text for prediction**

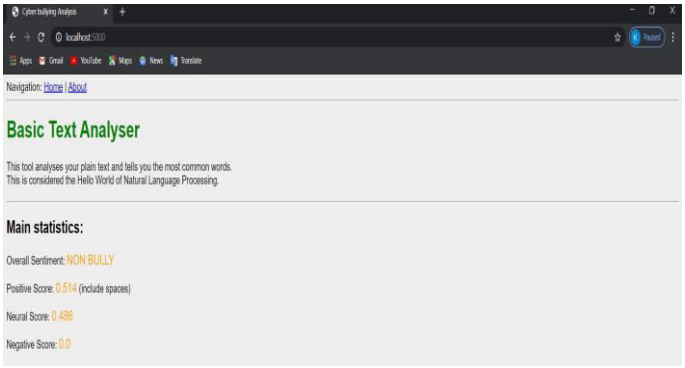

Fig 3:- basic text analyszer

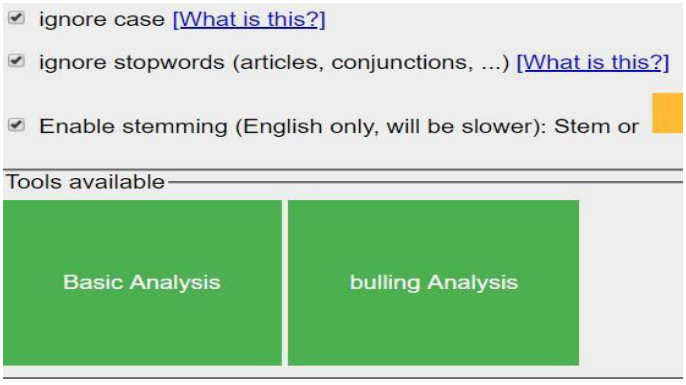

Fig 4:- bullying Analyzer

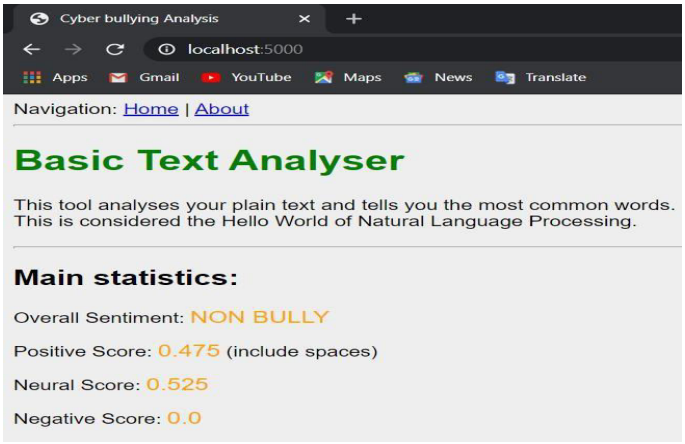

Fig 5: **:- classification of bully and non bully**

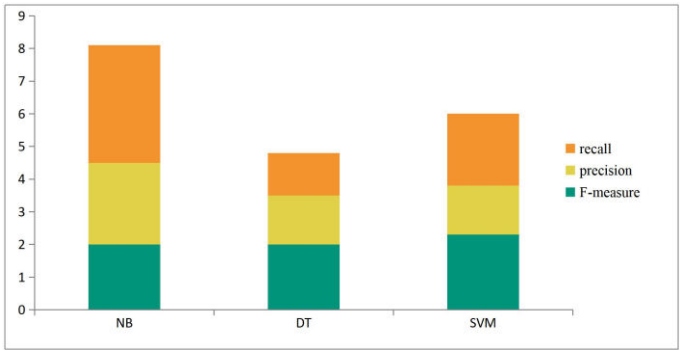

fig 6:- evolution metrics accuracy of machine learning models

Query Text	Num True Positives	TP in top 10	Precision
Q: whooorrreeeeee A: No liifffeeee	1	1	1.00
Q: hoe A: thankssssssssssssssss	99	9	0.85
Q: shove it/ A: shove what bitch	260	9	0.79
Q: NIGGER A: sup bitch	260	9	0.78
Q: your a hoe A: anonomous	104	9	0.77
Q: **bitch*** lol VVV A: .	268	9	0.76
Q: bitch you kant rapp niqquh.! A: suk my dik pussy	321	9	0.76
Total	1313		

fig 7:- analysis of query tweets

6 Conclusion

We hope to improve automated identification of cyberbullying as part of this study, which is a first step toward automated systems for monitoring modern social situations that can have a negative impact on mental health. We created a framework for detecting harassment-based cyberbullying with minimal supervision, which eliminates the requirement for human specialists to undertake time-consuming data annotation. Language describing subpopulations of different social groupings should be treated equally by an ideal, fair language-based detector. To identify serious online aberrations and prevent them from spreading, early detection of detrimental social media behaviours such as cyberbullying is required. However, automated identification is just one of several issues that must be addressed in order to fully manage the cyberbullying epidemic. This study is a significant step forward in terms of technology capability for detecting cyberbullying automatically.

7 References

1) Y. Altshuler, M. Fire, E. Shmueli, Y. Elovici, A. Bruckstein, A. S. Pentland, and D. Lazer. The social amplifier - reaction of human communities to emergencies. Journal of Statistical Physics, 152(3):399–418,2013.

2) N. V. Chawla. Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook, pages 853–867. Springer,2005.

3) D. C. Campfield. Cyber bullying and victimization: Psychosocial characteristics of bullies, victims, and bully/victims. ProQuest,2008.

4) M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong. Improving cyberbullying detection with user context. In Advances in Information Retrieval, pages 693–696. Springer,2013.

5) T. D´avid-Barrett and R. Dunbar. Processing power limits social group size: computational evidence for the cognitive costs of sociality. Proceedings of the Royal Society B: Biological Sciences, 280(1765),2013.

6) K. Dinakar, R. Reichart, and H. Lieberman.

Modeling the detection of textual cyberbullying. In The Social Mobile Web,2011.

7) V. Nahar, X. Li, and C. Pang. An effective approach for cyberbullying detection. Communications in Information Science and Management Engineering, 3(5):238–247,2013.

8) K. Reynolds, A. Kontostathis, and L. Edwards. Using machine learning to detect cyberbullying. In Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on, volume 2, pages 241–244. IEEE,2011.

9) Hinduja, S. and J.W. Patchin, 2008. Cyberbullying: An exploratory analysis of factors related to offending and victimization. Deviant Behav., 29: 129-156. DOI: 10.1080/01639620701457816.

10) Juvonen, J. and E.F. Gross, 2008. Extending the school grounds?--bullying experiences in cyberspace. J. Schools Health, 78: 496-505. PMID:18786042.

11) Kowalski, R.M. and S.P. Limber, 2007. Electronic bullying among middle school students. J. Adolesc. Health, 41: 22-30.DOI:10.1016/j.jadohealth.2007.08.017.

12) Li, Q., 2006. Cyberbullying in schools a research of gender differences. School Psychol. Int., 27: 157-170. DOI:10.1177/0143034306064547.

13) Li, Q., 2007. New bottle but old wine: A research of cyberbullying in schools. Comput. Hum. Behav., 23: 1777-1791. DOI:10.1016/j.chb.2005.10.005.

14) Luan, W.S., N. Siew F. and H. Atan, 2008. Gender Differences in the usage and attitudes toward the internet among student teachers in a public Malaysian University. Am. J. Applied Sci., 5: 689-697. DOI:10.3844/.2008.689.697.

15) Mason, K.L., 2008. Cyberbullying: A preliminary assessment for school personnel. Psychol. Schools, 45: 323-348. DOI:10.1002/pits.20301.

16) Slonje, R. and P.K. Smith, 2008. Cyberbullying: Another main type of bullying? Scand. J. Psychol., 49: 147-154. PMID:18352984