IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Recognize and Prevent the Cyberbullying Conversation on Social Networks Using Machine Learning Techniques

Dr.K.N.S LAKSHMI [#1], MADUTURI NIKITHA [#2]

[#1] Professor, Department of Computer Science and Engineering,
Sanketika Vidhya Parishad Engineering College, P.M. Palem,
Visakhapatnam, Andhra Pradesh.

[#2] MCA Student, Department of Computer Science and Application,
Sanketika Vidhya Parishad Engineering College, P.M. Palem,
Visakhapatnam, Andhra Pradesh.

ABSTRACT

Social media was a common place for abusive communication in the modern day. More than 80% of online social networks have abusive or vulgar speech on their user profiles, according to a recent survey report. Cyberbullying is the act of threatening or harassing another user via the use of inappropriate, abusive, or vulgar online social media posts. The major purpose of putting these kinds of unpleasant statements on user walls is to harass teenagers, preteens, and other kids. I created the present programme to limit vulgar communication on online social media since, up until now, no application has been able to stop this cyber material from spreading in social networks online. In this suggested application, our major goal is to provide a novel representation learning approach to address the issue of recognising and preventing abusive statements in online chat. Here, we attempt to categorise the misused and legitimate text messages using well-known machine learning methods, including Porter Stemming Algorithm and Support Vector Machine (SVM). The NLT Package (Natural Language Toolkit), which divides the entire message into pieces and then assigns tokens for each and every unique word, is known as Porter Stemming. Here, we divide online bullying into five types, including "HATE, VULGAR, OFFENSIVE, SEX, and VIOLENCE," that are found in the literature.

KEYWORDS:

Support Vector Machine, Natural Language Toolkit, Vulgar, Porter Stemming Algorithm

1. INTRODUCTION

Social media, as used here, refers to a collection of web-based tools that support the production and sharing of user-generated content and expand on the philosophical and technical principles of Web 2.0. People may take use of a wealth of knowledge, easy communication, and other benefits through social media. Social media could have unfavourable effects on people's lives, particularly those of children and teenagers, such as cyberbullying. Because they do not have to confront anyone and can hide behind the internet, bullies are free to damage their friends' sentiments. Because we are all always linked to the Internet or social media, especially young people, victims are readily exposed to harassment.

According to [2], victimization rates for cyberbullying range from 10% to 40%. Nearly 43% of teens in the US have experienced cyberbullying at some point [3]. Cyberbullying has same detrimental, sneaky, and pervasive effects on kids as conventional bullying does [4], [5], and [6]. The results of cyberbullying for victims may even be tragic, like the occurrence of self-harming behaviour or suicides.

Automated detection and fast reporting of bullying communications is one strategy for combating the issue, allowing for the right action to be done averting potential catastrophes. Natural language processing and machine learning are effective techniques for studying bullying, according to earlier computational research [7, 8]. Detecting cyberbullying may be described as supervised learning problem.

The numerical representation for Internet messages should be accurate and discriminatory in cyberbullying detection. To lessen uncertainty, strong representations of these communications are needed because social media posts are sometimes very brief and full of casual language and typos. Even worse, the problem is made more difficult by the scarcity of enough high-quality training data, or data sparsity. First of all, classifying data takes a lot of time and effort. Second, because of its inherent ambiguities, cyberbullying is challenging to define and assess from a third-party perspective. Thirdly, the majority of posts on bullying are removed from the Internet owing to concerns about privacy and the protection of Internet users. The trained classifier may not generalise effectively as a result.

The objective of the current work is to create techniques for detecting cyberbullying that can acquire robust and discriminative representations in order to address the aforementioned issues. Expert knowledge has been included into feature learning in certain methods to address these issues[10]. To expand the general features, Dynacare g. Yin et. al suggested to integrate BoW features, sentiment features, and contextual features. The label specific features were learnt using Linear Discriminative Analysis [11] and were used to extend the general features. Furthermore, common sense information was used. Scaling bullying-like traits by a factor of two resulted in Nahar et a presentation [12].

2. LITERATURE SURVEY

Literature survey is that the most vital step in software development process. Before developing the new application or model, it's necessary to work out the time factor, economy and company strength. Once all these factors are confirmed and got an approval then we can start building the application.

MOTIVATION

According to Kaplan and Haenlein [1], the idea of social media is currently at the top of the agenda for many corporate leaders. Both consultants and decision-makers look for ways that businesses might profit from using platforms like Twitter, Facebook, Second Life, YouTube, and Wikipedia. The purpose of this page is to give some explanation because, despite this interest, there appears to be very little knowledge of what the word "Social Media" actually implies. Start with defining the term "Social Media" and comparing it to terms like "Web 2.0" and "User Generated Content" to show how it differs[13]. Using this definition as a foundation, present a classification of social media that divides the applications currently grouped under the umbrella term into more precise categories according to a characteristic: collaborative projects, blogs, content communities, social networking sites, virtual game worlds, and virtual social worlds. Finally, offer 10 pieces of guidance for businesses considering using social media.

Cyberbullying, according to Ybarra[2], is the use of technology as a tool to bully someone. Bullies have a lot of room to manoeuvre on social networking sites, which makes teenagers and young adults who use them vulnerable to assault. The linguistic patterns that bullies and their victims employ have been identified using machine learning, and criteria have been developed to automatically identify content that is bullying others online.

Latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data like text corpora, is described by Blei et al. in their discussion of "Latent Dirichlet Allocation" in the Journal of Machine Learning Research (see reference 3). Each item of a collection is described as a finite mixture over an underlying set of topics in the three-level hierarchical Bayesian LDA model. The model for each subject is an infinite mixture over a base set of topic probabilities. The topic probabilities in the context of text modelling offer an explicit representation of a document. Describe effective approximation inference methods for estimating the empirical Bayes parameters based on variation approaches and an EM algorithm.

Discussed by Kontostathis et al. in section 4 Teenage bullying is a significant public health concern. Social media presents a fresh chance to examine bullying in both the real world and online. With the use of sentiment analysis, bullying may be better understood scientifically and victims who represent a high danger to themselves or others may be identified. Name seven feelings that are prevalent in bullying. While certain emotions have been well researched in the past, others are not commonly seen in the sentiment analysis literature. Provide a quick training method to identify these emotions without intentionally creating a typical labelled training dataset. Apply our process to online bullying posts, and then talk about your results. Discovered a wide spectrum of emotions in bullying traces and suggested a quick learning process to teach a model to recognise them automatically.

3. EXISTING SYSTEM AND ITS LIMITATIONS

There was no pre-defined technique or programme in the current system to categorise the abusive or cyberbullying messages for a text message posted on OSN walls, recognise the meaning of that word, and stop that message from being placed directly on the user's wall.

Therefore, the following are the limits of the current system. They are listed below.

LIMITATION OF PRIMITIVE SYSTEM

The following are the limitations of the existing system.

1) Up till now, there hasn't been a method like SEMdae in the literature to automatically identify and encode cyberbullying communications.

2) No phrase like BoW exists in the current system, which uses a bag of words that are entered into a database to match the dimensions of a term that is placed on the wall.

4. PROPOSED SYSTEM AND ITS ADVANTAGES

We exploited expert knowledge for feature learning in the suggested system. In order to train a support vector machine for online harassment detection, the proposed system leverages ML-Approach to categorise the semantic meanings of uploaded messages. We also attempt to incorporate BoW features, sentiment features, and contextual features.

The following are the advantages of our proposed system.

1) Most cyberbullying detection methods rely on the BoW model.
2) This should be verified or managed by the Administrator while adding words into the BoW database.
3) In this proposed application by giving labels for the BoW, we can get an exact count of each and every word like how many abused words are used in the message and which word come from which category.

5. IMPLEMENTATION PHASE

The step of implementation is when the theoretical design is translated into a programmatically-based approach. The application will be divided into a number of components at this point and then programmed for deployment.

JSP, HTML, and Java Beans are used for the application's front end, and My SQL was used for the back end database. The following five modules make up the bulk of the application. These are what they are:

1. Network Construction Module
2. Marginalized Stacked Denoising Auto-encoder Module
3. Semantic Enhancement for mSDA Module
4. Construction of Bullying Feature Set Module
5. Label Feature Selection Module

Now let us discuss about each and every module in detail as follows:

1) Network Construction Module

In this module initially we need to construct a network containing single admin and multiple users. Where the admin has the facility to add a set of words into each BoW database based on individual category. The admin should add each and every word into the database individually. Once if a word is added in one category the same word shouldn't be added on another category. So this should be mandatory step for the admin while adding words into the database. Also admin has the facility to authorize each and every user at the time of registration. The user who got activated by admin only can access his profile by login into the site. Those users who are not authorized can't be enter into their individual accounts at any cost[14].

2) Marginalized Stacked Denoising Auto-Encoder Module

In this module we has the facility to identify each and every word in a message and the words which are matched with BoW will be automatically identified and recognized as noise word and that will be automatically encoded into the application they were treated as a cyber bulling message[15].The basic idea behind denoising auto-encoder is to reconstruct the original input from a corrupted one ~x1,…., ~xn with the goal of obtaining robust representation.

3) Semantic Enhancement for mSDA Module

Here the semantic enhancement module is nothing but earlier algorithms used to identify exact word from a sentence which is matched from the Bow. But they failed to identify the related words which are nearer to the cyber bulled word. But in this application we can able to identify the words that are similar and related to the BoW. So this is treated as a semantic enhancement of mSDA.

4) Construction of Bullying Feature Set Module

The bullying features play an important role and should be chosen properly. In the following, the steps for constructing bullying feature set Zb are given, in which the first layer and the other layers are addressed separately. For the first layer, expert knowledge and word embeddings are used. For the other layers, discriminative feature selection is conducted.

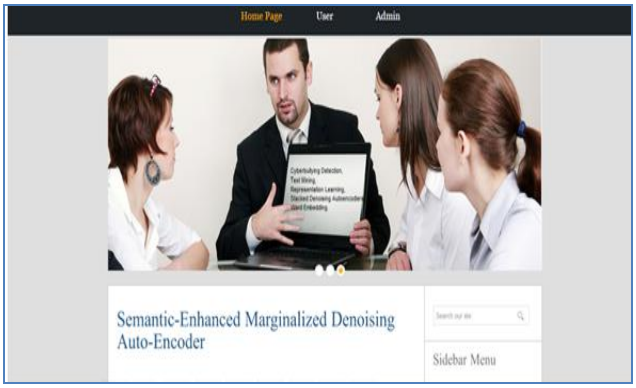
5) Label Feature Selection Module

Here we proposed a labeled Feature Selection method where the labeling is done because ,if any word is matched from a set of BoW,then they are automatically identified as a abused word and they will be identified based on individual category wise.Hence labeled based feature selection method is mainly used for categorizing each and every matched word based on category wise.

6. EXPERIMENTAL RESULTS

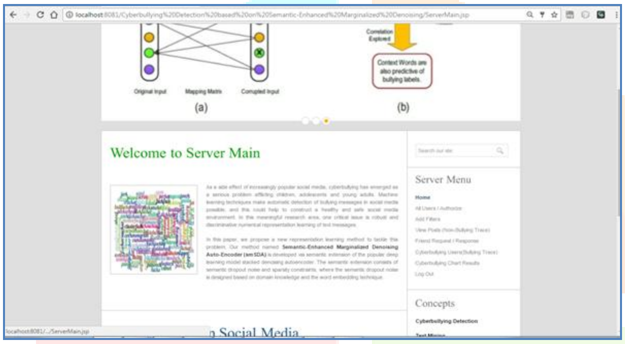
In this section we try to design our current model using Java as programming language and we used J2EE as working environment for executing the application and MYSQL as backend database for showing the performance of our proposed application. The front end of the application we use JSP,HTML and CSS and as a back end we used my-sql as database for showing the performance of our proposed application.Now we can check the performance of our proposed application as follows:

MAIN WINDOW



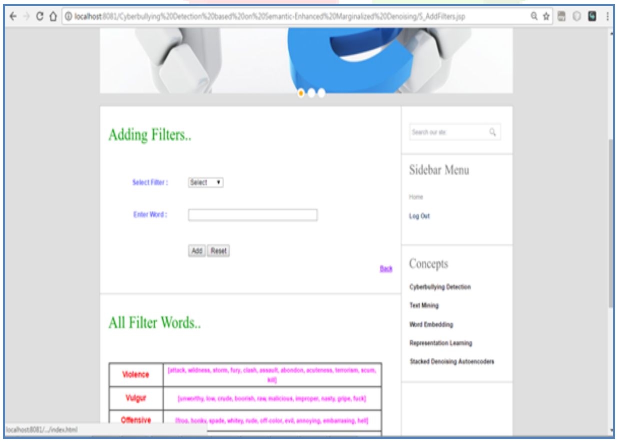
The above window clearly represent the main window.

ADMIN MAIN PAGE


localhost:8081/Cyberbullying%20Detection%20based%20on%20Semantic-Enhanced%20Marginalized%20Denoising/ServerMain.jsp
Original Input
Mapping Matrix
Corrupted Input
(a)
Correlation Explored
Context Words are also predictive of bullying labels
(b)
Welcome to Server Main
Search our site
Server Menu
Home
All Users / Authorize
Add Filters
View Posts (Non-Bullying Traces)
Friend Request / Response
Cyberbullying Users(Bullying Traces)
Cyberbullying Chart Results
Log Out
Concepts
Cyberbullying Detection
Text Mining
Social Media
localhost:8081/.../ServerMain.jsp

From the above window, we can clearly identify list of facilities available in the admin.

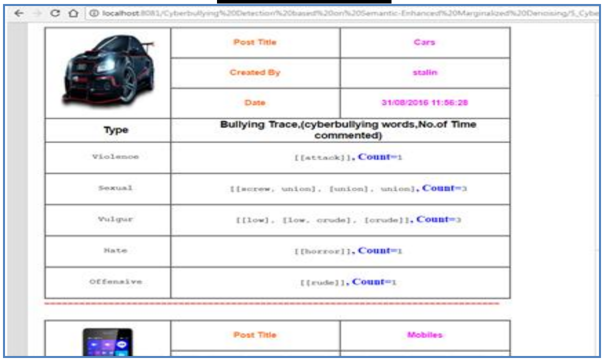
ADMIN ADD WORDS INTO THE DATABASE



Violence	[attack, wildness, storm, fury, clash, assault, abondon, acuteness, terrorism, scum, kill]
Vulgur	[unworthy, low, crude, boorish, raw, malicious, improper, nasty, gripe, fuck]
Offensive	[froo, honky, spade, whitey, rude, off-color, evil, annoying, embarrasing, hell]

From the above window we can clearly identify admin can add set of cyberbulled or abused words into the database and then try to maintain a Bag of Words(BoW) for classification of normal and cyber bulled messages from the social conversations.

ADMIN CAN SEE LIST OF CYBER MESSAGES


localhost:8081/Cyberbullying%20Detection%20based%20on%20Semantic-Enhanced%20Marginalized%20Denoising/S_Cybe

	Post Title	Cars
	Created By	stalin
	Date	31/08/2016 11:56:28
Type	Bullying Trace,(cyberbullying words,No.of Time commented)	
Violence	[[attack]], Count=1	
Sexual	[[screw, union], [union], union], Count=3	
Vulgur	[[low], [low, crude], [crude]], Count=3	
Hate	[[horror]], Count=1	
Offensive	[[rude]], Count=1	

	Post Title	Mobiles

From the above window, we can clearly identify the list of cyber bulled and normal messages. The messages can able separated with separate labels.

7. CONCLUSION

As a specific representation learning model for cyberbullying detection, we created a unique semantic-enhanced marginalisation de-noising auto encoder for the first time in this study. By creating the suggested technique, we may identify the individuals conducting abusive and poor communication inside a network and delete the abused messages from transmission. By taking into account word order in messages and employing natural language processing techniques to predict any abusive words that are not in the dataset and include the same by way of feedback into the BoW database, it is possible to further increase the robustness of the learnt representation.

8. REFERENCES

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of social media," Business horizons, vol. 53, no. 1, pp. 59–68, 2010.

[2] R. M. Kowalski, G. W. Giumetti, A. N. Schroeder, and M. R. Lattanner, "Bullying in the digital age: A critical review and metaanalysis of cyberbullying research among youth." 2014.

[3] M. Ybarra, "Trends in technology-based sexual and non-sexual aggression over time and linkages to nontechnology aggression," National Summit on Interpersonal Violence and Abuse Across the Lifespan: Forging a Shared Agenda, 2010.

[4] B. K. Biggs, J. M. Nelson, and M. L. Sampilo, "Peer relations in the anxiety–depression link: Test of a mediation model," Anxiety, Stress, & Coping, vol. 23, no. 4, pp. 431–447, 2010.

[5] S. R. Jimerson, S. M. Swearer, and D. L. Espelage, Handbook of bullying in schools: An international perspective. Routledge/Taylor & Francis Group, 2010.

[6] G. Gini and T. Pozzoli, "Association between bullying and psychosomatic problems: A meta-analysis," Pediatrics, vol. 123, no. 3,pp. 1059–1065, 2009.

[7] A. Kontostathis, L. Edwards, and A. Leatherman, "Text mining and cybercrime," Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK, 2010.

[8] J.-M. Xu, K.-S. Jun, X. Zhu, and A. Bellmore, "Learning from bullying traces in social media," in Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, 2012, pp. 656–666.

[9] Q. Huang, V. K. Singh, and P. K. Atrey, "Cyber bullying detection using social and textual analysis," in Proceedings of the 3rd International Workshop on Socially-Aware Multimedia. ACM, 2014, pp. 3–6.

[10] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, "Detection of harassment on web 2.0," Proceedings of the Content Analysis in the WEB, vol. 2, pp. 1–7, 2009.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying." in The Social Mobile Web, 2011.

[12] V. Nahar, X. Li, and C. Pang, "An effective approach for cyberbullying detection," Communications in Information Science and Management Engineering, 2012.

[13] M. Dadvar, F. de Jong, R. Ordelman, and R. Trieschnigg, "Improved cyberbullying detection using gender information," in Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop(DIR2012). Ghent, Belgium: ACM, 2012.

[14] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in Advances in Information Retrieval. Springer, 2013, pp. 693–696.

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," The Journal of Machine Learning Research, vol. 11, pp. 3371–3408, 2010.

About the Authors

Dr.K.N.S LAKSHMI is currently working as a Professor in Department of Computer Science and Engineering at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh. She has more than 16 years of teaching experience. Her research interest includes Machine Learning, Adhoc Networks, Network Security, and Python.

MADUTURI NIKITHA is currently pursuing her 2 years MCA in Department of Computer Science and Applications at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh.Her area of interest includes Python, Java, C, and C++.