

NLP using Wordnet Database for Monitor the Social Platform in Digital Forensic Investigation

P. Senthil¹, S. Alagumuthukrishnan² and K. S. Kannan³

¹CMR College of Engineering & Technology /Department of Computer Science and Engineering, Kandlakoya, Medchal Road, Hyderabad, India -501401
Email: sssenthil.p@gmail.com

²CMR Institute of Technology / Department of Computer Science and Engineering, Kandlakoya, Medchal Road, Hyderabad, India -501401

³Kalasalingam Academy of Research and Education / Department of Computer Science and Engineering, Krishnankoil, Tamilnadu, India - 626126
Email: alagumuthukrishnan@gmail.com, saikannan2012@gmail.com

Abstract—Cyberbullying has made people too vulnerable to hatred and abuse which makes cyberbullying an important subject of matter all over the world. We propose a framework that prevents the use from using abusive words and predict user crime probability in social application with the help of one of the machine learning technique Natural Language Processing (NLP). NLP plays an important role in analyzing the given words and find out the alternative meaning and provide which category these words belongs to with the help of the WordNet database. In the previous system, social application user information is processed and kept aside and required is action not taken. We proposed an application that solves that existing system issue and provides a solution by making a complaint on the user in the social application. The proposed system consists of three application social book, social admin, crime reporting. First social book is social media application that provides activities like adding friends posting status and etc. Second social admin has the control of the whole social media application admin has all details of the user and so on Third crime application contains two-part public and police this application acts as a compliant system to our application where public user can raise a complaint against a person in social application to take some action.

Index Terms— word net; Natural Language Processing; social platform; crime; cyberbullying; Cybercrime; complaint.

I. INTRODUCTION

Cybercrime or computer-oriented crime is a crime that involves computers and the internet. Cybercrimes can be defined as “Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or mental harm, or loss, to the victim directly or indirectly, using modern networks such as internet and mobile phones. One of the major cybercrime encountered is cyberbullying.

Cyberbullying is defined as the use of electronic communication to bully a person, typically by sending messages or comments of an intimidating or threatening nature. Harmful bullying behavior can include posting rumors, threats, sexual remarks, a victims' personal information, or pejorative labels (i.e. hate speech). Bullying

or harassment can be identified by repeated behavior and an intent to harm. Victims of cyberbullying may experience lower self-esteem, increased suicidal ideation, and a variety of negative emotional responses including being scared, frustrated, angry, or depressed. Victims of cyberbullying may not know the identity of their bully, or why the bully is targeting them, based on the online nature of the interaction. The harassment can have wide-reaching effects on the victim, as the content used to harass the victim can be spread and shared easily among many people and often remains accessible long after the initial incident. Nowadays cyberbullying has made people too vulnerable to hatred and abuse which makes cyberbullying an important subject of matter all over the world. The major function of social networking sites is to connect people and organizations from all over the world. Social media has provided a very effective medium for people to exchange their ideas and thoughts to millions of people in a single go. But not all the people use it for constructive purposes; some exploit these platforms for their sadist pleasures. Cyber bullying has a huge effect on the mind of the victim.

The main intention of this research is to find any type of abusive comments on the social platform and report the case to the respected official for further actions to be taken. Here the victim is free to lodge a complaint; hassle-free and can get a solution immediately. As the system admin has information of all the accounts on the platform, necessary implementation of action and further information about the offender is easily accessible.

II. DESIGN IDEAS AND RELEVANT CONCEPTS

A. *Design ideas*

The main aim of this system is to analyze the user activities to predict the probability of crime and make an online complaint to the police in case of offence. The proposed system consists of three application social book, social admin, and Crime reporting application:

1-Social book: It is the social media application that provides activities like adding friends posting status and etc.

2-Social admin: Admin has the control of the whole social media application admin has the all details of the user and so on.

3-Crime reporting application: It is divided into two part public and police this application acts as a compliant system to our application where public user can raise a complaint against a person in social application to take some action.

Crime reporting application acts as an Investigation system it contains two parts:

- Public
- Police.

Public are the general users of social book application .User has to register themselves in the crime reporting application with her or his credentials and login using valid credentials. After that in public part user login in that UI(user interface) ant it shows details like Online complaint status, status of CIR, status of FIR, Trace location and See criminal icon to check if there is any criminal present in the current location then by clicking the online complaints icon user can file a complaint about another person.

Likewise police also has to register her or his credentials and login with valid credentials. Similar to the public UI the Police UI (user interface) contains received complaints from the user and icons like Crime rate, cybercrime reporting, criminal record, criminal investigation.

B. *Design ideas*

In Crime reporting application public user can raise a complaint against a social user after that user login by clicking the complaint icon user need to fill following fields like complaint name, evidence, location etc this complaint form is forwarded to the police it is police choice to accept the complaint or not by clicking the accept button the complaint is taken by the police and accepted status will be notified to the public user when user checks his CIR status icon then by clicking cybercrime bar in Crime reporting application in police part. In cybercrime part police fills details of the social user details to request user chat data from social admin then police apply Natural language processing to existing abusive data which police already got then admin forward the requested user chat and details of the user to the police after that downloading the chat data and applying NLP after that comparing already existing abusive chat data which police got with admin sent user chat data and finding the probability of the Crime.

C. *NLP WordNet*

The aim of processing natural language is to create and synthesize computer systems that can interpret, understand and synthesize natural human languages. In the aim of understanding and generating meaningful phrases in the human language, natural language falls within the scope of artificial intelligence.

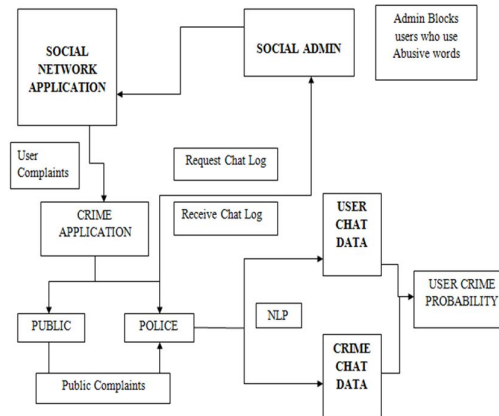


Figure 1. NLP Architecture for crime investigation

WordNet is a massive and publicly available, English language lexical database that was developed at Princeton University. It contains more than 120,000 concepts including nine noun hierarchies (80,000 concepts) and 554 verb hierarchies (13,500 concepts)[1]. WordNet is primarily designed for applications such as text mining and artificial intelligence. In contrast to a conventional dictionary containing word spelling, pronunciation, synonyms, and antonyms, WordNet describes the semantic association between word senses. A synset (synonym set) that reflects a distinct definition forms WordNet's fundamental building block. Specific language objects, i.e. noun, verb, adjective, and adverb, are organized into networks of synsets that are interrelated semantically and linguistically[11].

In the semantic network [13] of WordNet, nouns and verbs are interrelated in a hierarchy without breaching the boundary of the part-of-speech. The relationship of subordination between lexicalized concepts in the hierarchy is called hyponymy. For example, spoon is a hyponym of cutlery. Hypernyms are abstract concepts stored at the higher level, and the more relevant/customized concepts, labeled hyponyms, are located at the lower level in the hierarchy. For example, color is a hypernym of red. WordNet discusses the interaction between other language artefacts, such as adverbs and adjectives, in addition to the is - an association between verbs and nouns. Meronymy is used as an important aspect of a relationship. A meronym denotes a word or other element that together with other elements constitutes a whole. Thus, 'bark,' 'leaf,' and 'branch' are meronyms of the holonym 'tree.'

III. METHODOLOGY

This section will discuss the components utilized for the system to efficiently run the system.

JDK is an important component of the system. The Java Development Kit (JDK) is software used for development environment used for developing Java applications and applets. It includes the Java Runtime Environment (JRE), an interpreter/loader (Java), a compiler (javac), an archiver (jar), a documentation generator (Javadoc) and other tools needed in Java development.

Apache Tomcat is an open source web server that is developed by Apache software foundation. It basically makes our Java Web applications to run on host and server based system and it is configured on local host port 8080. It generally runs JSP, Servlet etc.

Client Server database is accessed in the google webpage. All the data are directed to MySQL databases and are stored in the form of corresponding directories inside a MySQL DATADIR directory, which is specified in a configuration. E.g. myExampleDB's files would be stored inside '\$DATADIR/myExampleDB' directory. And according to this result, database files would be stored inside '/var/db/mysql/%DB_NAME%' directory

This section will use NLP WordNet for processing the data from the user complain and the information available with the social admin.

WordNet can make connections among the words/concepts that humans often don't make.

Following methodologies is used in WordNet to check similarity among the given set of information.

A. Leacock Chodorow Similarity

Measure is a lexical semantic similarity measure based on (Resnik, 1995)'s information content.

$$Sim(W_i, W_j) = Max \left[-\log \frac{Dist(c_i, c_j)}{2xD} \right]$$

$$Sim(W_i, W_j) = Max \left[\log 2D - \log Dist(c_i, c_j) \right],$$

where $Dist(c_i, c_j)$ is the shortest distance between concepts c_i and c_j .

- Length is measured in nodes traversed (not edges); synonyms have length 1
- D = maximum depth of hierarchy (assuming unique beginner)

B. Wu and Palmer (1994) Conceptual similarity

$$Sim_{wup}(C1, C2) = \frac{2*N}{N1+N2+2*N}$$

- $N1$ = length of path $c1$ to lowest common subsume (LCS) of $c1, c2$
- $N2$ = path length $c2$ to $c3$
- $N3$ = path length $c3$ to root node Greater distance of nodes to LCS => lower similarity

C. Information-based similarity measure (Resnik)

This measure uses the information content of the shared parents. The principle of this measure is as follows: two concepts are more similar if they present a more shared information, and the information shared by two concepts $C1$ and $c2$ is indicated by the information content of the concepts that subsume them in the taxonomy. Resnik measure is formally defined as follows:

$$Sim_{Resnik}(C1, C2) = -\ln(p_{mis}(C1, C2))$$

- Thus, the higher the LCS of a pair of concepts, the lower their similarity (it's 0 for the root node)

D. X-Similarity

In 2006, Petrakis et al. proposed a feature-based function called X-similarity that proposes matching words retrieved from WordNet by parsing meanings of terms. Two terms are comparable when the conceptions of the words and the meanings in their neighborhoods are lexically equivalent (based on semantime relations).

$$Sim_{xsim}(A, B) = \begin{cases} 1, & \text{if } S_{synsets}(A, B) > 0 \\ \max\{S_{neighb}(A, B), S_{descr}(A, B)\}, & \text{if } S_{synsets}(A, B) = 0 \end{cases}$$

Where A and B denote a set of synsets or definitions for words a and b , the similitude for S_{descr} definitions and synonyms $S_{synsets}$ shall be determined as follows:

$$S(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

IV. RELATED WORK

WordNet can be used to analyze cybercrime using chat information and identification of criminal groups is possible by clique detection[1]. WordNet can be effectively used to improve retrieval performance without the requirement of very accurate query keywords to retrieve the most relevant context [7]. Natural language processing techniques make it possible to automatically identify offensive messages in social media, and this may help promote a healthy and secure social media ecosystem [15].

In crime analysis, the main responsibility for the development of the concept is to establish reasonable linkages in unstructured crime data. Unstructured crime reports can be used to extract plausible associations. Through text mining [3]. Criminal profiling is a beneficial method for determining the most likely culprits based on the evidence identified at the scene of the crime. A thorough review of all factors related to cybercrime will provide the investigator with clear evidence to trace the attacks and narrow the reach of the criminal data and eventually apprehend the criminals[2]. Content features, sentiment features, and contextual features of documents can also be utilized for detecting harassment and abuse over a social platform. Even as bullying is a significant problem in online communities, there are still no known approaches for automatically and effectively detecting online harassment [16].

Writeprints procedure can also be utilized for recognizing anonymous identities and discovering connections. Writeprints is a process built on Karhunen-Loeve transformations, which uses a sliding window and pattern disruption algorithm with individual processor architectures at the author level [11].

Online radicalization and civil unrest are two important application of Intelligence and Security Informatics (ISI) [6].

An individual's probability value is a defendant that can be estimated by considering the correlation between the terminology that can be extracted from the SEMCON as well as the criminal ontology principles. Web crawlers are useful for acquiring and processing user data like those of blogs, feeds, comments, etc. linguistically and conceptually employing objective metric ontology development SEMCON [8]. Computational linguistics in natural language processing and psycholinguistics can be used to identify cybercriminal-to-cybercriminal and second type is cybercriminal-to-victim communications and help the targeted victim or the community [13][4].

K-means clustering with certain modifications help in the process of crime pattern recognition.

Clustering algorithm for a data mining approach can also be used to detect the crimes patterns and speed up the process of solving crime [14]. Apache spark can be used for analyzing the Big Data generated by various social platforms to identify the real approved account, taking into consideration various requirements [6]. The classifier Naïve Bayes is applied to identify comments as abusive or not abusive. And efficiency can be accessed via the use of 10-fold cross-validation on minimally processed results [17].

A. Social Book Application

In this module the end user has to register with the Social book where the user has to give their basic details like name, address, mobile number, email id etc. The user has to enter valid user name and password checks validation finally user interface will be displayed to the user. Social book application contains activity like adding friends, posting status, commenting status of other friends, recommending friends to the user etc.

Position figures and tables at the tops and bottoms of pages. Table I is centre aligned, table II is left aligned and table III is right aligned. Avoid placing tables and figures in the middle of a page. Leave sufficient room between the figures/tables and the main text. Figure captions should be centered below the figures; table captions should be centered above. Avoid placing figures and tables before their first mention in the text. Use the abbreviation “Fig. 1, Fig. 2 and Fig. 3” even at the beginning of a sentence. All figures and tables must cite in the text using figure and table number.

To figure axis labels, use words rather than symbols. Do not label axes only with units. Do not label axes with a ratio of quantities and units. Figure labels should be legible, about 9-point type.

Color figures will be appearing only in online publication. All figures will be black and white graphs in print publication. Screenshot, programs, and other junk images etc must be removed; instead block diagrams, algorithm, flowchart, circuit diagram, state diagram, UML designs, etc may be introduced.

The paper must have proposed system, results, discussion to infer the quality of the research paper. All the figures, equations and etc must be in high resolution and in good quality.



Figure 2. Social book application

B. Social Book Admin

In this module Social book admin controls the Social book application admin has the access privilege to access user comments logs in the application. Admin has two responsibility first is to add spam keywords and upload

image that indicate some kind of abusive content these spam keywords and image is used to block users who uses words and image that are similar to spam keywords and image in the application before that these spam keywords are processed using natural language processing Social admin UI (user interface) contains social users and police icons. In social users part admin can view list of social users and in police part admin can view the request came from police to get the details of the users and chat data.

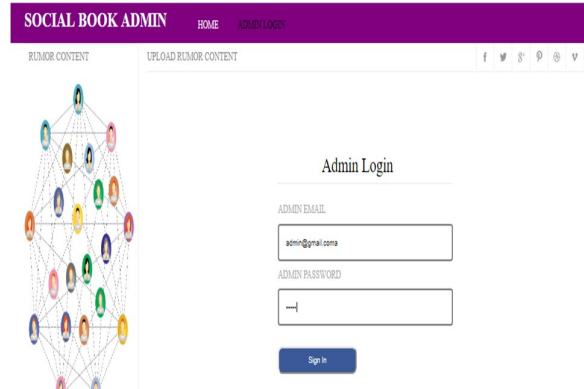


Figure 3. Social Book Admin login

C. Crime Reporting Application

Crime reporting application acts as an Investigation system it contains two parts Public and Police. Public in the sense are the general users of social book application or any people User has to do registration in Crime reporting application with her or his credentials and login using valid credentials likewise police also has to register her or his credentials and login with valid credentials after that In public part user login in that UI(user interface) it shows details like Online complaint status, status of CIR, status of FIR, Trace location and See criminal icon to check if there is any criminal present in the current location then by clicking the online complaints icon user can file a complaint about another person same as in Police UI(user interface) contains received complaints from the user and icons like Crime rate, cybercrime, criminal record, criminal investigation.

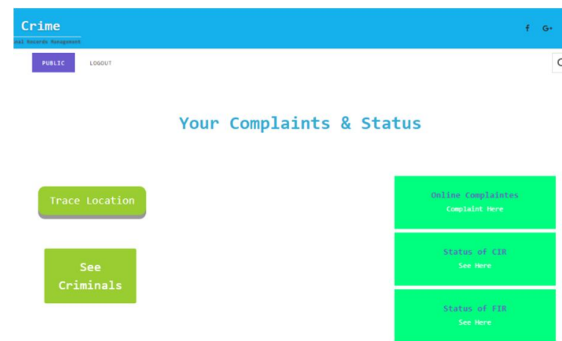


Figure 4. Current status for raised complaints

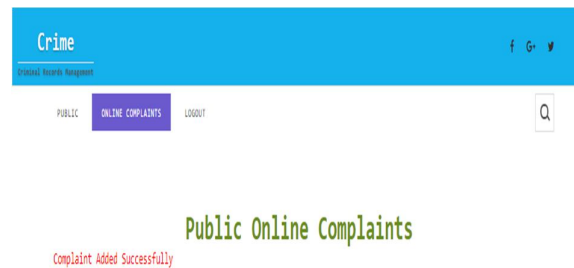
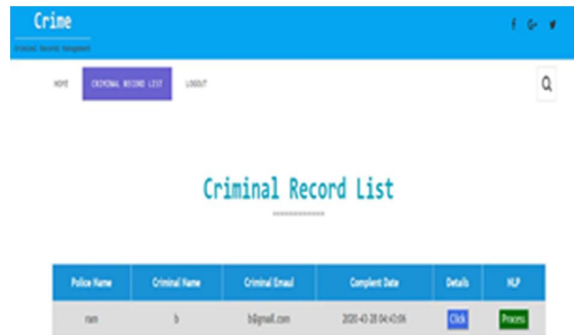


Figure 5. Crime Reporting Application

D. Investigation

In Crime reporting application public user can raise a complaint against a social user after that user login by clicking the complaint icon user need to fill following fields like complaint name, evidence, location etc this complaint form is forwarded to the police it is police choice to accept the complaint or not by clicking the accept button the complaint is taken by the police and accepted status will be notified to the public user when user checks his CIR status icon then by clicking cybercrime bar in Crime reporting application in police part. In cybercrime part police fills details of the social user details to request user chat data from social admin then police apply Natural language processing to existing abusive data which police already got then admin forward the requested user chat and details of the user to the police after that downloading the chat data and applying NLP after that comparing already existing abusive chat data which police got with admin sent user chat data and finding the probability of the Crime.



The screenshot shows the 'Crime' header with a navigation bar containing 'HOME', 'CRIMINAL RECORD LIST', and 'LOGOUT'. Below the header is a search bar. The main content area is titled 'Criminal Record List' and displays a table with the following data:

Police Name	Criminal Name	Criminal Email	Complaint Date	Details	NLP
ram	b	b@gmail.com	2020-07-28 04:42:08	Click	Process

Figure 6. Criminal record list analyzation



Figure 7. Blocking



Figure 8. NLP

V. CONCLUSION

A social platform is a web-based technology that enables the development, deployment and management of social media solutions and services. It provides the ability to create social media websites and services with complete social media network functionality. Everyone on the platform is vulnerable to some sort of abuse or hatred. This paper introduces an a framework that prevents user from using abusive words and predict user Crime probability in social application with the help of one of the machine learning technique Natural Language Processing (NLP).NLP plays important role in analyzing the given words and find out the alternative meaning and provide which category these words belongs to with the help of Word Net database. In previous system social application user chat logs are processed and kept aside and required action not taken by keeping these chat data we cannot find the user future crime level activity to overcome this problem We proposed an application that solve that existing system issue and provide a solution by making a complaint on user in social application. In the future, planning to incorporate a system that will enable the user to directly contact the authority and seek remedies, then also to implement an online counseling facility for the affected which can be accessed from anywhere in the world.

REFERENCES

- [1] J. Lu, G. Tang, H. Zhang, P. Jiang, W. Wu, "A Review of Research Methods for Spatiotemporal Distribution of the Crime Hot pots," *Progress in Geography*, 2012, vol.31, no.4, pp.419-425.
- [2] H. Liu, D.E. Brown, "Criminal incident prediction using a point-patternbased density model.," *International Journal of Forecasting*, 2003, vol.19, no.4, pp.603-622.
- [3] Dragolju Pokrajac, Reed L. Hoskinson and Zoran Obradovic, "Modeling Spatial-Temporal Data with a Short Observation Histroy," *Knowledge and Information System*, 2003, vol.5, no.3, pp.368-386.
- [4] Ivaha, Hasan Al-Madfai, Gary Higgs and Andrew Ware, "The Dynamic Spatial Disaggregation Approach: A Spatio-Temporal Modelling of Crime," In *Proceedings of the Word Congress on Engineering* 2007, London, U.K, July, 2007, vol.2.
- [5] A. Tayebi, Martin Ester, Uwe Glasser, Patricia L. Brantingham, "Crime Tracer: Activity Space Based Crime Location Prediction," *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, August, 2014, pp.472-480.
- [6] [6] Ieivan Kianmehr, Reda Alhajj, "Crime Hot-Spots Prediction Using Support Vector Machine.," *IEEE International Conference on Computer Systems and Applications*, March, 2006, pp.952-959.
- [7] H. Liu, D.E. Brown, "Criminal incident prediction using a point-patternbased density model.," *International Journal of Forecasting*, 2003, vol.19, no.4, pp.603-622.
- [8] Donald E. Brown, Hua Liu, Yifei Xue, "Mining Preferences from Spatial-Temporal Data," In *Proceedings of the 2001 SIAM International Conference on Data Mining*, 2001, pp.1-17.
- [9] T.H. Grubestic, E.A. Mack, "Spatial-temporal interaction of urban crime," *Journal of Quantitative Criminology*, 2008, vol.24, no.3, pp.285-306.
- [10] I. Jayaweera, C. Sajeewa, S. Liyanage, and T. Wijewardane, "Crime analytics: Analysis of crimes through newspaper articles," *Moratuwa Engineering Research Conference (MERCon)*, IEEE, April 2015, pp. 277 – 282.
- [11] R. Arulanandam, B.T.R. Savarimuthu, and M.R. Purvis, "Extracting crime information from online newspaper Articles," *Proceedings of the Second Australasian WebConference*, 2014, Vol.155, pp. 31 – 38.
- [12] K.R. Rahem, and N. Omar, "Drug-Related crime information extraction and analysis," *International Conference on Information Technology and Multimedia (ICIMU)*, Malaysia, IEEE, November 2014, pp. 250 – 254.
- [13] Jayaweera, C. Sajeewa, S. Liyanage, and T. Wijewardane, "Crime analytics: Analysis of crimes through newspaper articles," *Moratuwa Engineering Research Conference (MERCon)*, IEEE, April 2015, pp. 277 – 282.
- [14] R. Arulanandam, B.T.R. Savarimuthu, and M.R. Purvis, "Extracting crime information from online newspaper Articles," *Proceedings of the Second Australasian WebConference*, 2014, Vol.155, pp. 31 – 38.
- [15] K.R. Rahem, and N. Omar, "Drug-Related crime information extraction and analysis," *International Conference on Information Technology and Multimedia (ICIMU)*, Malaysia, IEEE, November 2014, pp. 250 – 254.
- [16] S.J. Conlon, A.S. Abrahams, and L.L. Simmons, "Terrorism information extraction from online reports," *Journal of Computer Information Systems*, Springer, 2015, Vol.55, Issue 3, pp. 20 – 28.