DETECTING THE CYBERBULLYING TWEETS USING MULTINOMIAL NAÏVE BAYES CLASSIFIER

Banoth Geethanjali[#1], Kamatam Sahasra[#2], Kodumuru Sankeerthana[#3], M.L.Aishwarya[#4] and chittakula Rohini [*5]

[#] *Research Scholar, Information Technology Department, VNR Vignana Jyothi Institute of Engineering and Technology,Hyderabad,India*

[*] *Research Guide, Assistant Professor,Information Technology Department, VNR Vignana Jyothi Institute of Engineering and Technology,Hyderabad,India*

***Abstract*— Increasing the use of Internet and facilitating access to online communities such as social media have led to the emergence of cybercrime. Cyberbullying is highly prevailing in this social environment that lowered the efficiency of the learning system. The detection of cyberbullying and online harassment is often formulated as a classification problem. Techniques typically used for document classification, topic detection, and sentiment analysis can be used to detect electronic bullying using characteristics of messages, senders, and the recipients. In this paper, a Multinomial Naïve Bayes Classifier is employed to detect the cyberbulling tweets with the help of sentimental analyzer tool. Initially, the dataset is collected from the public repositories and thus, it is preprocessed using TF-IDF technique. The words and their topics are trained under the scope of multinomial NB classifier. With the help of cyberbullying trained model, the testing tweets are classified into two classes, namely cyberbullying tweets and the normal tweets. The proposed classifier is executed on Python programming language that helps to easily design an efficient detection system. The analytics have proved the efficiency of the proposed classifier by classifying the 84000 records with a stipulated period of time.**

***Index Terms*— Twitter; Cyberbullying; Naïve Bayes classifier; TF-IDF and sentimental analysis.**

I. INTRODUCTION

Dissemination and detection of cyberbullying tweets of social media have become a major challenge for the past several years. Due to the popularity gained by social media, millions of data are generated on a daily basis [1]. Likewise, different devices are also associated with social media that led to an increased presence of data. Different social platforms like Twitter, YouTube, Facebook and so on will act as great information source to the media content developers. It shares the real-time data which helped for promoting the contents. These promotions may be spread as false tweets, irrelevant words, etc at some time constraints. Henceforth, social media has become a powerful tool for different sorts of journalism like sports, healthcare and politicals and also encouraging cyberbullying activities [2].

A. Source of tweets:

Since social media content is used, the content (or) data may be in the form of text, video and audio. Each type of data spread false information in four forms, namely [3],

Rumours: It is a piece of information where the information is insufficient with fewer trust factors. Mostly, it is created by panic towards public opinions of different credibility.

Fake news: In order to gain the attention of the readers, the false news are written and published.

Misinformation: Some inappropriate information is communicated as if it is true content.

Hoax: Most of the political news was hoaxed by journalists, so as to gain the attention of the news.

B. Types of tweets:

The general categorization of the tweets is classified as follows [4]:

Visual based features: It includes graphical representation of the fake news like images, video and so on.

User-based features: Several fake accounts were created by social users. Thus, the target variables are included in age, groups etc.

Post-based features: It appears in social media contents like tweets, memes etc.

Network-based features: It describes the associativity between users and networks.

Propagation-based features: It describes how false sentences are written and published.

Knowledge-based features: It provides false content of unresolved issues.

C. Sentimental Analysis:

Propagation-based features are prominent study among the researchers. The opinions of different users provoke the emotional-aspects of the human. Therefore, the detection of cyberbullying-related tweets has to be addressed predominantly. Rapid developments made in information processing system has helped for user-generated content from different sources like blogs, social media etc., that covers both academia and industrial aspects of opinions and sentiments. Here, aspects based opinion mining was used for topic modeling system [5]. By doing so, we shall design an efficient class model and thus understanding the aspect-based sentiment analysis. The art of dealing with opinions of generated test is known as sentiment analysis. It has greatly developed interest among the researchers and academia for last decades. It is a kind of technique that also helps for finding the hidden patterns and its context. The application of topic modeling is in processing large amounts of non-formalized text with possible datasets using different correlation studies. The algorithms applied in topic modelling include tokenization, phrase detection, lemmatization and word. With the help of these techniques of topic modelling, more information and useful measures can be created. In specific to, depending on the user-provided reviews, percepting its information type is a complex and challenging task. In recent times, it is followed by rating systems, highest rated information are used for modelling the topics which helps for future prediction such as cost, efficiency and customer services. Some sequences of topics are irrelevant to the product which is known as unsupervised learning process. At last, different forms of opinion are extracted from reviews on different entities.

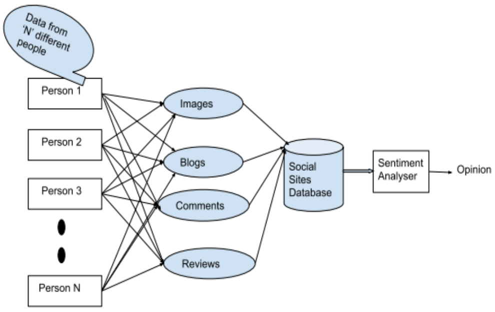


Fig.1. Theoretical view of social media

D. *Organization of this study:*

The paper is organized as follows:

Section 2 presents the Related Work that explains the reviews of existing techniques explored to overcome the challenges of the detecting cyberbullying in twitter.

Section 3 presents the Proposed Work that discusses the novel technique designed to enhance the performance of detection system.

Section 4 presents the Experimental Results and Discussion that explores the validation procedures followed to prove the designed detection system's efficiency.

Section 5 presents the Conclusion that dictates the findings of the proposed detection system.

II. Related Work

This section presents the reviews of existing sentimental analyser techniques employed for the proposed study. Topic modelling is the process of extracting the relevant topic from its semantic data structure with the help of corpus. It provides different insights on a given texts that leads to higher computational overload. It fetches the texts and search by specified keywords, so as to acquire its relevant topics [6]. Consider an instance, topic is the domain, and thus, domain of mobile, LCD and monitor differ to picture and screen of the movie domain. It differs from its aspects. Henceforth, topic modelling has been extensively focussed by the researchers, so as to enhance the web contents. One of the techniques was Bayesian modeling for hierarchical viewpoint discovery from tweets for progressive perspective revelation [7] from tweets. By joining the earlier data into the model, it can be observed that a noteworthy lift in position order precision where pHOP accomplishes an exactness of 71% and beats sLDA by 2%. HASM just accomplishes a precision of 60% since it just uses the conclusion vocabulary and can't profit by the intermediary position names because of its unbounded angle hub course of action for each sentence, which needs to pursue rCRP.

Opinion mining based on the extracted features has been carried out in many studies. In an intelligent analysis and summarization models to review the data was suggested in [8]. Performing summarization on distinguishing emotions are not easy to interpret by the users. Thus, a feature based opinion mining was suggested for easier interpretation of results. The machine learning algorithms for sentiments process was discussed in [9]. As we know, bag-of- words (BOW) is commonly used for representing the analysed sentiments. Semantics and independency of the words helps for obtaining subjective information presented in the texts. High dimensionality in feature space develops higher computational complexity during topic modelling. Then, authors [10] suggested new semantics model that combines the natural language processing and sentiment analysis process. The system focussed on ontology which helped for selecting the papers and the vector analysis approach was also used for mining the sentiments. In the realistic view, the obtained results were efficient, but the time complexity of task scheduling is higher than the conventional techniques.

To reduce this complexity the binary classification schemes to enhance the sentimental intensity was proposed in [11]. It was combined with the SVM using 10-fold cross validation. Here, hybrid PSO was also employed to enhance the accuracy by improving the local search and its capabilities. In view of ambiguities, lexical analysis was further improved by [12]. Along with that, context aware modeling was also analysed for polarity checking, in order to derive with knowledge base. An intelligent feature selection process that combined PSO and Conditional Random Field (CRF) learning framework was developed and was experimented on SemEval-2014 datasets and achieved F-measure 81% and 72% in laptop and restaurant domain [13]. Then, the study was extended by [14] using Genetic Algorithm (GA) and the RST function based on NLPs. Subset candidates and its pruning lowers the performance of the training data, which is being resolved by performing meta-heuristic search models. It not only prunes the subsets but also dictates what sorts of features helpful for designing sentimental classifier. Yet, the system failed to remove the redundant features.

In [15], the authors presented an extraction sentiment model that have utilized the POS based rule and the dependency relations of given document. Later on, phrase orientation was done for each class which helped for aggregating polarity of documents. In [16], the authors devised a prototype for sentiment classification of movie reviews with Efficient Repetitive Pre-processing (SentReP) which depends on tested parameter and focused pre-processing technique for classifying the opinion. This method works on the Cornell movie review data sets and proves the accuracy and efficiency of SentReP along several volumes of data compared to other existing approaches and hence considered as efficient technique for sentiment analysis.

III. Proposed Work

This section presents the proposed technique designed to enhance the performance of the detection system in twitter using classification approach. In this study, Naïve Bayes (NB) classifier is proposed that deliberately enhances the efficiency of the cyberbullying tweets with the help of sentimental analyser components. The fig.2 explains the system architecture of this study wherein the efficiacy of supervised classifier is modulated.

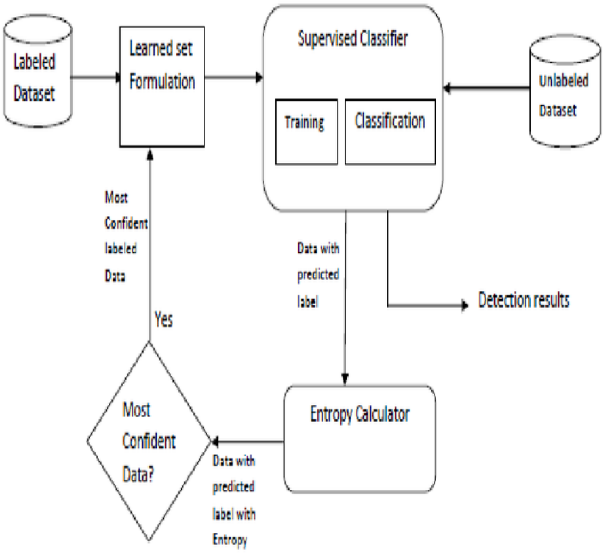


Fig.2 System Architecture

A. Data collection & Preprocessing:

Data collection is the first step of this study. The tweets are collected from the Kaggle, public repositories. Since the data is collected from the open environment, it is being subjected to hold uncertain data. It is processed using preprocessing technique. The main task of this stage is to clean and label the collected data via preprocessing techniques. Its main intention is to eliminate the stopwords which are not relevant in the documents. Examples, pronouns, suffix, prefix and the similar roots are eliminated in this step. By doing so, it helps to reduce the high data dimensionality prevails in the collected documents. Term Frequency_ Inverse Document Frequency (TF-IDF) is one of the efficient stemming techniques employed to preprocess the collected data. Since it is independent towards the language, diverse form of n-gram stemming approach is done. Most Arabic words are obtained from the three-letter root word, 3 –gram stemming approach is performed. It estimates the similarity between words using Dice's Coefficient (DC), which are given as,

$$DC = \frac{2C}{A+B} \qquad (1)$$

Where,

A → Count of unique digrams in the first word.
B → Count of unique digrams in the second word.
C→ Count of unique digrams shared by A and B

The obtained DC values are considered to construct the similarity matrix for all words in the document. It is represented as,

Document = {word $_1$; word $_2$ ... word $_{n-1}$}

Similarity matrix:

Word$_1$
Word$_2$ s21
Word $_3$ s31 s32
....
....
Word$_n$ sn1 sn2 sn3 sn(n-1)

B. Feature Selection

This is the most important stage in which the obtained output from this step determines the efficiency of the building classifiers. Prior steps help to transform the unstructured text into the structured format. By using the structured data, the relevant features of the given document are extracted by applying the feature selection techniques. In alignment with the previous steps, we have taken the document features from the outcome of the 1-gram; 2-grams and 3-grams techniques. By estimating the weight of each word in the document, the words are categorized on the basis of highest weights.

C. Classification

Multinomial NB classifier is a simple classifier that performs bayes theorem by means of probabilistic distribution approach. To deal with the text classification problem, it is expressed as,

P (class C_i | documentD_i) = P (C_i). P (D_i| C_i) / P (D_i) (6)

Where,

P (class C_i | documentD_i) → the chance of document D belongs to the class C.

P (D_i) → the chance of a document

P (C_i) → It represents the number of classes presented in a document by exploring its categories.

P (D_i| C_i) → the chance of documents for the considered class.

Then, the class with the highest probability is obtained from bayes theorem, which is expressed as,

$$C^{"}(D_i) = \arg max_j P(C_i|D) \qquad (7)$$

The conditional probability of the data point d$_i$ of Document D is estimated as,

$$P(D|C_j) = \prod_i P\ (d_i\ |C_i) \qquad (8)$$

Then, the eqn. (6) combining with the above eqn. (8) can be rewritten as,

P (class C_i | documentD_i) = $\frac{P\ (C_i).\prod_i P\ (d_i|C_i)}{P\ (D_i)}$ (9)

Relied upon the classification problems, the classes are derived for the categorized documents. In this study, we have selected multi-variable Bernoulli NB model integrated with the Bayes theorem by assigning the value (+1) for the features related to the document, else it assign to the value (-1).

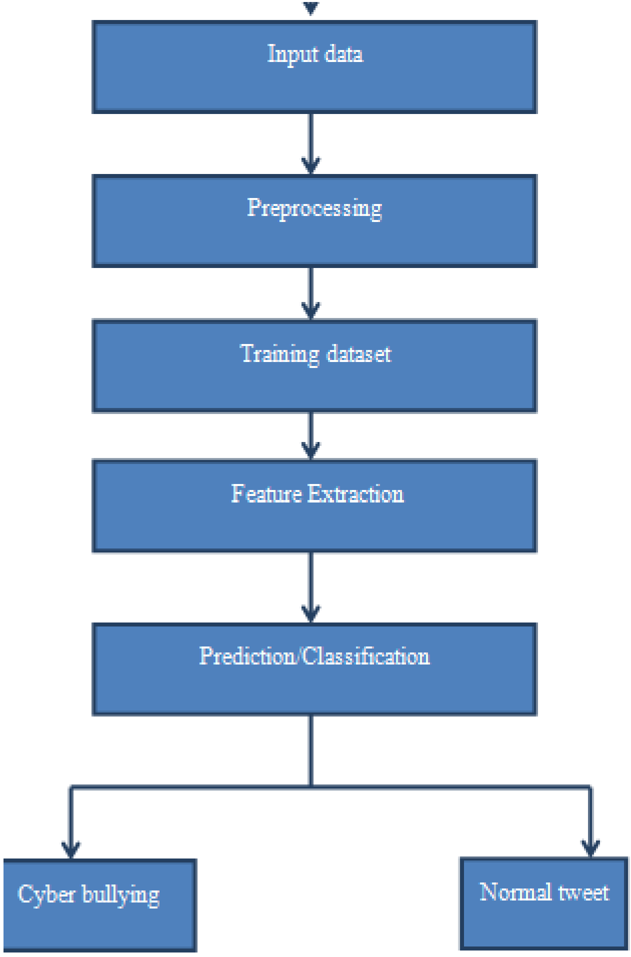
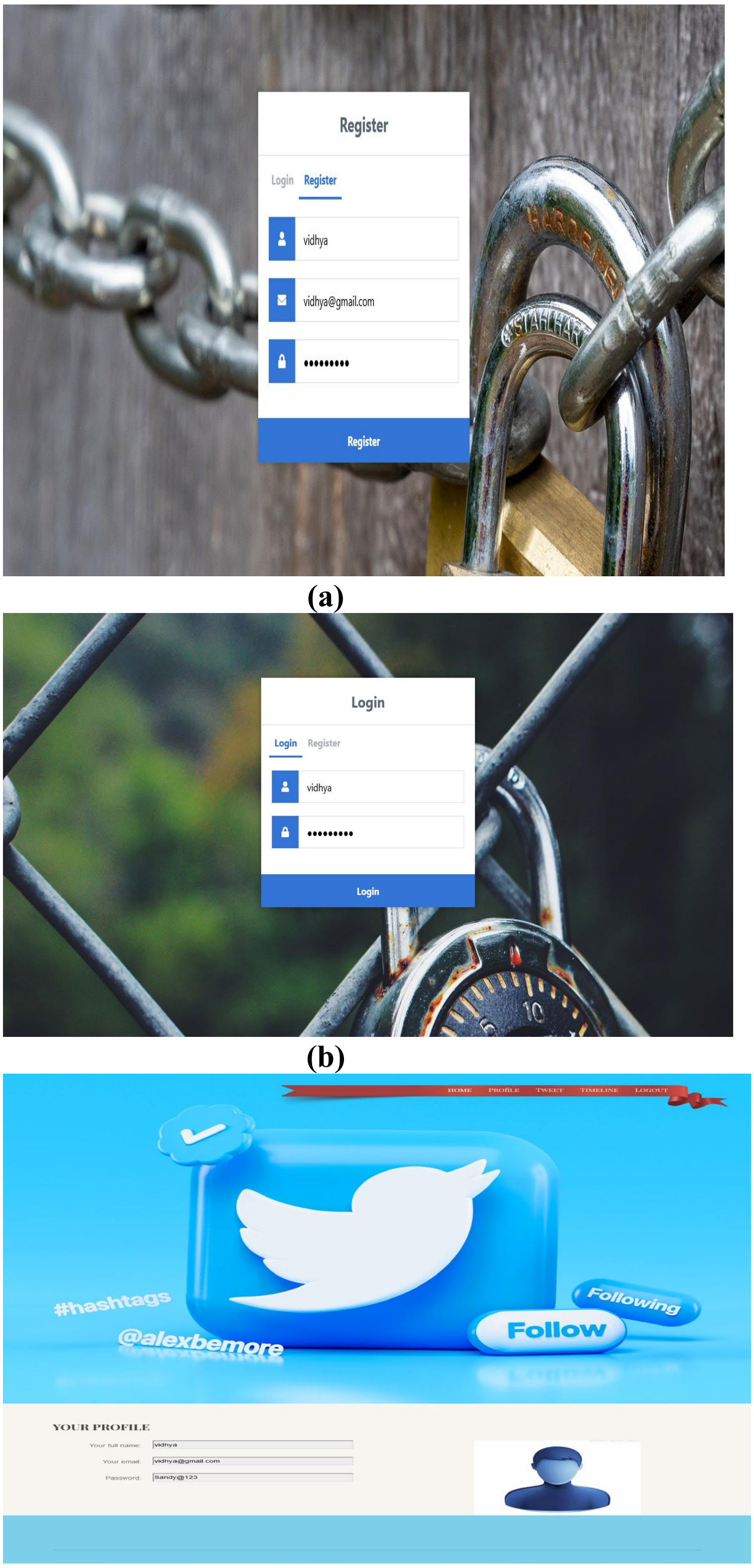


Fig. 3. Proposed Workflow

IV. Experimental Results and Discussion

This section presents the experimental setup and the achieved results. The proposed Naïve Bayes classifier is implemented in Python programming language with the backend support of MySQL. The dataset is collected from the [17] that contain 4 attributes, 2 classes and 85000 records. In our study, we have two attributes i.e tweets and its annotation. The 2 classes are cyberbullying tweet (1) and normal tweet (0).



vidhya
vidhya@gmail.com
•••••••••

(a)

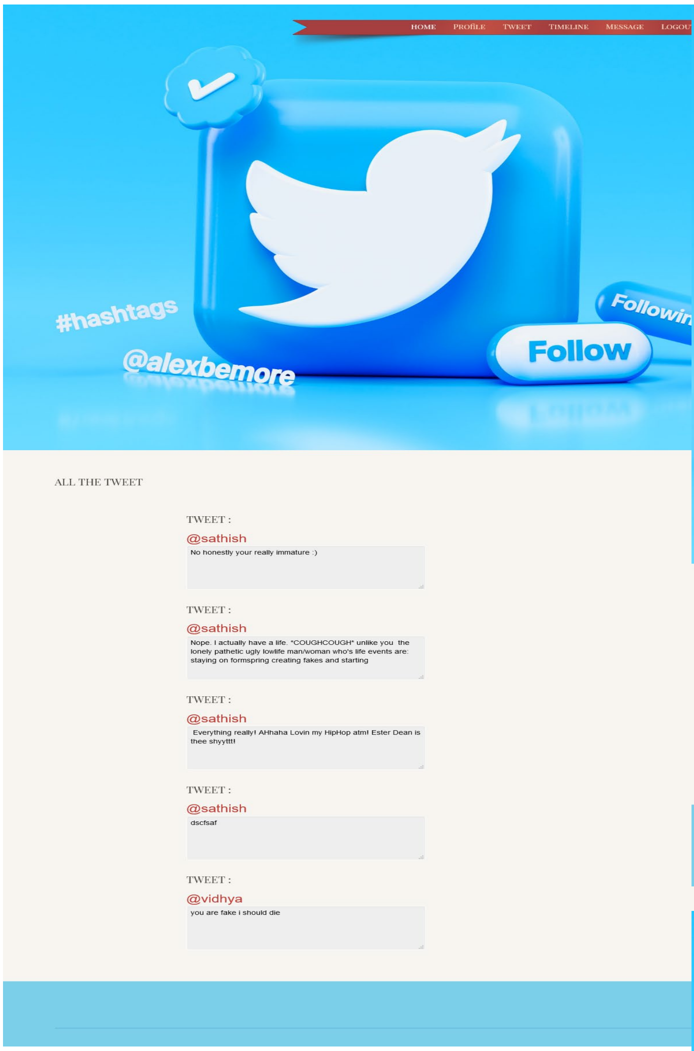
vidhya
•••••••••

(b)

Your full name: vidhya
Your email: vidhya@gmail.com
Password: Sandy@123

(c)

Fig. 4. (a-c) User’s registration, login and profiling process



ALL THE TWEET

TWEET :
@sathish
No honestly your really immature :)

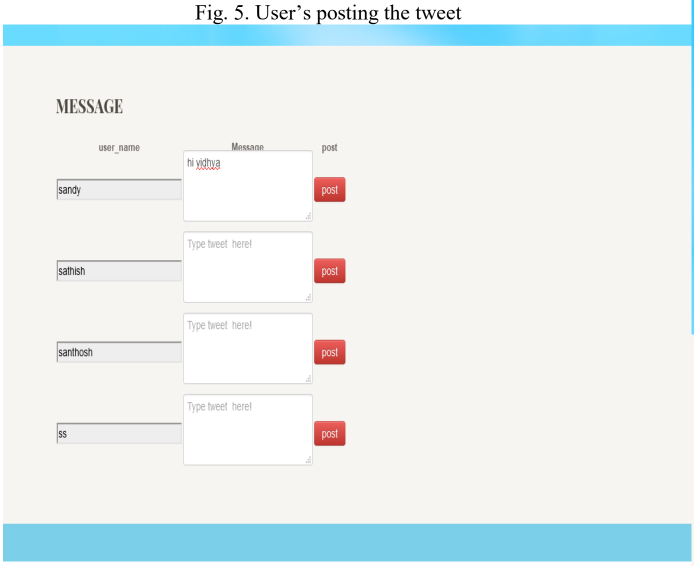
TWEET :
@sathish
Nope. I actually have a life. *COUGHCOUGH* unlike you the lonely pathetic ugly lowlife man/woman who's life events are: staying on formspring creating fakes and starting

TWEET :
@sathish
Everything really! AHhaha Lovin my HipHop atm! Ester Dean is thee shyyttt!

TWEET :
@sathish
dscfsaf

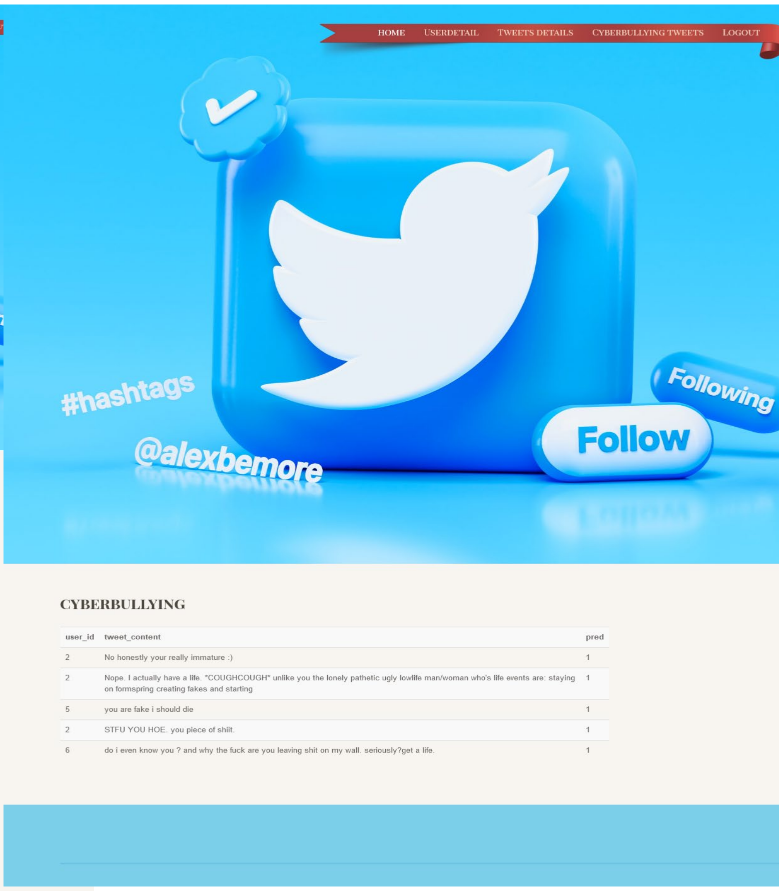
TWEET :
@vidhya
you are fake i should die

Fig. 5. User's posting the tweet


MESSAGE

user_name	Message	post
sandy	hi vidhya	post
sathish	Type tweet here!	post
santhosh	Type tweet here!	post
ss	Type tweet here!	post

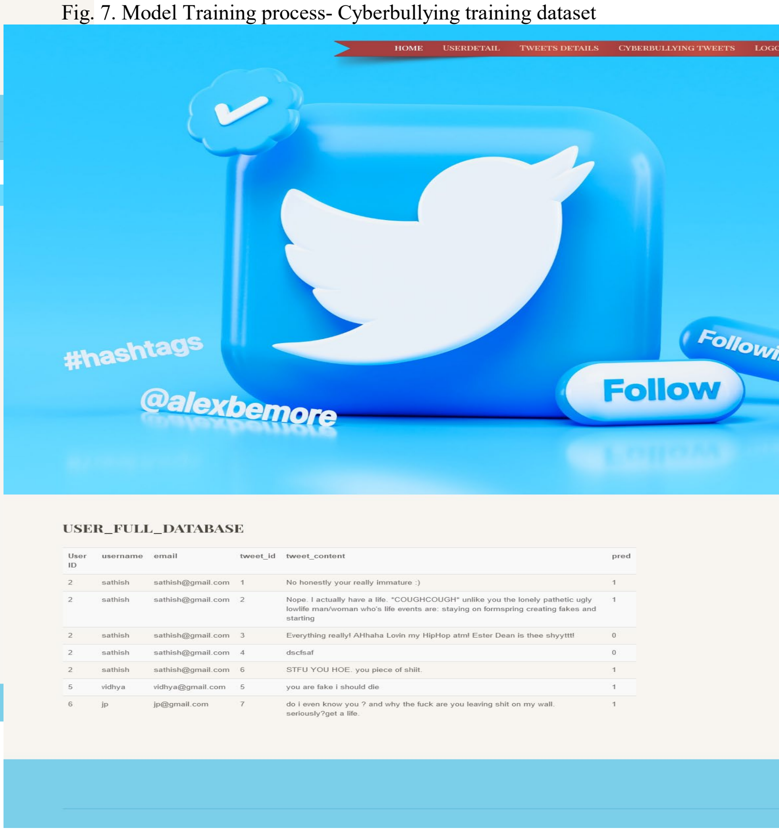
Fig.6. Preprocessing the tweet



CYBERBULLYING

user_id	tweet_content	pred
2	No honestly your really immature :)	1
2	Nope. I actually have a life. *COUGHCOUGH* unlike you the lonely pathetic ugly lowlife man/woman who's life events are: staying on formspring creating fakes and starting	1
5	you are fake i should die	1
2	STFU YOU HOE. you piece of shiit.	1
6	do i even know you ? and why the fuck are you leaving shit on my wall. seriously?get a life.	1

Fig. 7. Model Training process- Cyberbullying training dataset



USER_FULL_DATABASE

User ID	username	email	tweet_id	tweet_content	pred
2	sathish	sathish@gmail.com	1	No honestly your really immature :)	1
2	sathish	sathish@gmail.com	2	Nope. I actually have a life. *COUGHCOUGH* unlike you the lonely pathetic ugly lowlife man/woman who's life events are: staying on formspring creating fakes and starting	1
2	sathish	sathish@gmail.com	3	Everything really! AHhaha Lovin my HipHop atm! Ester Dean is thee shyyttt!	0
2	sathish	sathish@gmail.com	4	dscfsaf	0
2	sathish	sathish@gmail.com	6	STFU YOU HOE. you piece of shiit.	1
5	vidhya	vidhya@gmail.com	5	you are fake i should die	1
6	jp	jp@gmail.com	7	do i even know you ? and why the fuck are you leaving shit on my wall. seriously?get a life.	1

Fig.8. Detecting the cyberbullying and normal tweets classes

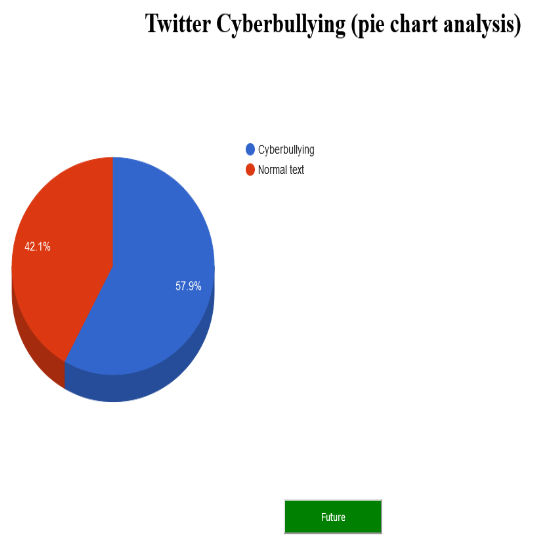


Fig. 9. Detecting the tweet classes for the entire records

V. Conclusion

Developments made in Information Processing System (IPS) have enabled the growth of marketing environment. In recent days, web environment comprises of different sorts of information. Due to higher invasion of dimensionality and the predictive parameters, the analysis on topic modeling is still in under- developmental stage. The primary aim of the sentiment analyser is to arrange the contents into positive, negative and neutral classes. Conventional techniques utilized rating system to classify the data. In recent studies, text forms as well as numeric form are examined to classify the data. Twitter is one of the eminent real-time applications wherein the cyberbullying activities are heavily contributed by the users. Therefore, the need of designing an efficient detection system is been addressed in this study using Multinomial Naïve Bayes classifier. The designed classifier has effects over the topics relating to the cyberbullying activities are trained and then classified with the help of components of sentimental analyzer tool. The proposed classifier has efficiently detected the normal and cyberbullying tweets at a faster pace. Experimental results have proven the efficacy of the detection system.

References

[1] M. S. Akhtar, D. Gupta, A. Ekbal, and P. Bhattacharyya, Feature selection and ensemble construction: a two-step method for aspect based sentiment analysis, Knowledge Based System, 125 ,2017, pp. 116–135.

[2] M. S. Akhtar, S. Kohail, A. Kumar, A. Ekbal and C. Biemann, Feature selection using multi-objective optimization for aspect based sentiment analysis, in: the International Conference on Applications of Natural Language to Information Systems, 2017, pp. 15–27.

[3] D. Hazarika, S. Poria, P. Vij, G. Krishnamurthy, E. Cambria and R. Zimmermann, Modeling inter-aspect dependencies for aspect-based sentiment analysis, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, 2, 2018, pp. 266–270.

[4] Abdul-Mageed, M., Diab, M., and Kubler, S.¨Samar: Subjectivity and sentiment analysis for arabic social media. Computer Speech & Language. 28 (1), 2014, pp. 20–37

[5] Agarwal, B., and Mittal, N. Machine learning approach for sentiment analysis. In Prominent feature extraction for sentiment analysis. Springer, 2016, pp. 21–45.

[6] Yang, M., Yin, W., Qu, Q., Tu, W., Shen, Y., & Chen, X. Neural Attentive Network for Cross-Domain Aspect-level Sentiment Classification. IEEE Transactions on Affective Computing. 2019

[7] Zhu, L., He, Y., & Zhou, D. Hierarchical viewpoint discovery from tweets using Bayesian modelling. Expert Systems with Applications, 116, 2019, pp. 430-438.

[8] Ganeshbhai, S. Y., & Shah, B. K. Feature based opinion mining: A survey. In Advance Computing Conference (IACC), 2015 IEEE International, 2015, pp. 919-923.

[9] Agarwal, B., & Mittal, N. Machine Learning Approach for Sentiment Analysis. In Prominent Feature Extraction for Sentiment Analysis, Springer International Publishing, 2016, pp. 21-45.

[10] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., &Hussain, A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation, 7(4), 2015, pp. 487-499.

[11] Basari, A. S. H., Hussin, B., Ananta, I. G. P., &Zeniarja, J. Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 2013, pp. 453-462.

[12] Weichselbraun, A., Gindl, S., & Scharl, A. Extracting and grounding contextaware sentiment lexicons. IEEE Intelligent Systems, 28(2), 2013, pp. 39-46.

[13] Gupta, D. K., Reddy, K. S., & Ekbal, A. PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis. In International Conference on Applications of Natural Language to Information Systems, 2015, 220-233.

[14] Ahmad, S. R., Bakar, A. A., &Yaakub, M. R. Metaheuristic algorithms for feature selection in sentiment analysis. In Science and Information Conference (SAI), 2015,pp. 222-226.

[15] Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., &Hussain, A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation, 7(4), 2015, pp. 487-499.

[16] Manek, A. S., Pallavi, R. P., Bhat, V. H., Shenoy, P. D., Mohan, M. C., Venugopal, K. R., & Patnaik, L. M. SentReP: Sentiment Classification of Movie Reviews using Efficient Repetitive Pre-Processing. In TENCON 2013-2013 IEEE Region 10 Conference (31194), 2013, pp. 1-5.

[17] https://md-datasets-public-files-prod.s3.eu-west-1.amazonaws.com/748c31a9-5291-413a-8123-91042c3fe72c