INTERNATIONAL JOURNAL · ENGINEERING RESEARCH ·

International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 15th Nov 2021. Link :http://www.ijiemr.org/downloads.php?vol=Volume-10&issue=ISSUE-11

DOI: 10.48047/IJIEMR/V10/I11/18

Title **ANALYSIS OF VARIOUS APPROACHES OF ABUSIVE TEXT DETECTION IN ONLINE SOCIAL NETWORKS**

Volume 10, Issue 11, Pages: 124-130

Paper Authors

Ch.Ravisheker, Dr. Manmohan Sharma

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

ANALYSIS OF VARIOUS APPROACHES OF ABUSIVE TEXTDETECTION IN ONLINE SOCIAL NETWORKS

[1] Ch.Ravisheker, [2]Dr. Manmohan Sharma
[1] Research scholar, Lovely professional University, Punjab, India
[2] Associate Professor, Lovely professional University, Punjab, India

Abstract

Social media is one of the most influential tool for sharing information across different regions among different users .The people sharing their interests in various aspects in online social networking platforms like Facebook, twitter etc. Therefore the usage of hate text steadily increasing. Nowadays it has been reviled unfair behavior of the users in social networking sites. The existence of abusive text on different online social networking platforms and identification of such text is a big challenging task. To understand the complexity of language constructs in different languages is very difficult .Already lot of research work has completed in English language. This paper gives detail analysis of detecting hate text in various languages Hindi, urdu, Arabic, Bengali, Telugu. We incorporated various kinds of ML and DL based algorithms to identify hate text in OSN's. A review is done related to different classifiers where a comparison made between different models of ML, DL algorithms. Finally finds the accurate method to classify the text is offensive or not by finding the parameters i.e. accuracy and F1score

Key Words: Abusive text, Online Social networks, NLP constructs, Machine learning, deep learning**.**

Introduction

Nowadays, OSN's becomes most popular and it becomes a part of day to day human life. Distinct social networks like YouTube, Facebook, Twitter, Instagram etc. The users are communicating with messages, comments etc. The users share their opinions about news, videos, and personals. Not all the information in social networking platforms are positive some of them is negative ones. According to scan safe's worldwide survey, around 80% of social media platforms contains harmful content. The people intended to hurt specific group or individual it is treated as hateful content. Eventually hate text can be outlined as the people comments on nationality, religion, shading, color, ethnicity, race, sex etc. In different regions different rules to accept the text is hate or not. Some countries states its parts of freedom of speech where as other regions comply strict rules for the same. Daily the usage of internet is increasing rapidly especially spreading abusive text in social networking sites. In 2018, Thailand, 21.6 million of users using Facebook.

According to the survey of Indonesian service user association in 2018, 143 millions of users are active in usage of internet in Indonesia. In social media the users can freely express their opinions, writing bad words, hurtful words. Currently some software's identifying hate and blocking their webpages. But these are not applicable for detecting hate text. In order to do opinion mining, classification ML algorithms like SVM, RF, NB, Decision tree are used. Moreover identifying of hate text matching with word to word is difficult task because each language has their own grammar structure. Nowadays OSN's plays a vital role in society. The people can share and express their thoughts freely. Due to the misuse of internet user may posts offensive and hateful comments. It is very challenging to identify such hate text manually by existing methods. Thus automatic identification of offensive text is required in OSN's. This paper gives detail analysis of various existing abusive text detection approaches and focused on parametric comparison of these techniques.

Existing Techniques

This part gives brief description of various hateful or offensive text detection techniques

Shahnoor C. Eshan et al. (Eshan & Hasan, 2018)

Shahnoor C.Eshan et al. implements(Eshan & Hasan, 2018) Machine learning techniques to detect abusive text in the Bengali language. The training data, as well as tested samples, are gathered from Bangladeshi Facebook celebrities' posts and evaluated the performance o(Eshan & Hasan, 2018)f ML algorithms. This paper focused classification of text by using different machine learning algorithms, SVM performs better than C4.5 and Naïve Bayes(Eshan & Hasan, 2018). One of the most common text classification problems is sentiment analysis. Sentiment analysis has been done on Bengali blog posts with a variety of feature extraction methods. SVM works with better performance with unigram and emoticons features. To filter the web content, applied Aho-corasick method which is based on searching string. Few experiments are conducted to detect abusive in Bengali language and tested ML methods.

Several experiments are carried out and posts are extracted from various Facebook celebrities in Bangladesh (Eshan & Hasan, 2018). All the lexemes '(', '-', '@' etc. are removed only Bengali comments with Unicode characters are considered. The comments are categorized as abusive or non-abusive. For validating results 90% trained data, 10% tested data used. Three types of word feature techniques unigram, bigram, and trigram applied to different types of strings(Eshan & Hasan, 2018). These features are used for training ML algorithms as well as vector creation, count vectorizer, and Tfid vectorized features. Experiments are conducted by taking 4 cases.

In the first case, unigram features combined with count vectorizer and acquired the results from machine learning algorithms(Eshan & Hasan, 2018). It is observed SVM with a linear kernel achieves best by measuring accuracy. The second case, compared RF, SVM, MNB algorithms with bigram count vectorizer features from these MNB has the best accuracy. Various ML algorithms compared with Tfidf vector and unigram features out all SVM along linear kernel achieves the best performance compared to bigram Tfid vectorized elements SVM with linear kernel performs good accuracy. After conducting experiments with trigram Tfidf vectorizer finally, it is noticed that SVM linear kernel classifier achieved good results. Moreover, Tf-IDF has more weight and less frequent important features(Eshan & Hasan, 2018). Also, MNB performs better in terms of accuracy with bigram and trigram features. In this paper Tfidf, vectorizer features with SVM linear kernel are more effective when compared to count vectorizer features. In the future more experiments to be conducted with neural network models like deep neural networks (DNN), LSTM, and CNN models

Discriminative multinomial naive Bayes algorithm (DMNB)(Tuarob & Mitrpanont, 2017)

Suppawong Tuarob et al. proposed a ML based method which automatically detect abusive language in social networks(Tuarob & Mitrpanont, 2017). This classifier detects the content from social media and shows abusive or not. In Thailand millions of the people are accessing social network from these majority is young children. The text classification problem is a challenging issue in Thai abusive language detection. To consign this issue, proposed Machine learning methods. To train the data various data preprocessing methods binary, TF, IDF and TF-IDF are explored. These trained data given to ML based classifiers. Nine classification algorithms are compared theses are NB, MNB, Max Entropy, SVM, and KNN, Decision table / Naïve Bayes hybrid, RF, RIPPER and C4.5 decision tree(Tuarob & Mitrpanont, 2017).

Dataset from Facebook API, comments selected from Facebook pages. Data is preprocessed by Thai word tokenizer (CTWT) classifier and calculated F-measure 93%. In this paper the abusive message categorized as vulgar, figurative, rude, and dirty messages. Experiments are conducted by comparing different classification algorithms with trained features weighting schemes. Among all DMNB with IDF feature weighted schemes performs best achieved F-measure of 86% on 3497 Tha**i** Facebook social community(Tuarob & Mitrpanont, 2017). In future work, the existing algorithms has to be test on different datasets. For improving the thai language efficiency need to explore more on grammatical structure and classification.

Devanagari Hindi Offensive Tweets (DHOT) Corpus(Jha et al., 2020)

Vikas kumar jha et al, proposed a model to label abusive text from devanagari hindi offensive tweets (DHOT) data(Jha et al., 2020). The proposed fast text model classifier which achieves 92.2% accuracy. According to boston survey 50% of people which are using internet from remote areas by 2020. They did common conversation on social networks mostly in variety of languages like Hindi, Punjabi, English, Tamil, etc. Data has collected from Twitter API with 150 hindi speaking people also build some abusive words in hindi. At first, data preprocessing has done on tweets and removed hash tags, URL, emoticons(Jha et al., 2020).

Second, for annotation of DHOT tweets, classify tweet either hate, abusive or humorous.

Third, implemented fast text, word2vec model to process data for experiments. In this paper most of the cases it is observed mixing of English letters written in Devanagari hindi. The main drawbacks are tagging data manually with specific class is time taking process. Lack of dictionaries and tracking protocols in these languages. Finally our proposed fast text model with ideal multicore CPU classified half million strings less than minute.(Jha et al., 2020)

Mohammad Pervez Akhter et al(Akhter et al., 2020)

Mohammad Pervez Akhter et al, proposed dataset of urdu and applied seventeen classifiers as well as seven ML approaches to identify offensive text from urdu and roman language(Akhter et al., 2020). Hindi and Urdu are very similar except the script. In this paper, the data has collected from YouTube as Urdu language and constructed dataset to detect offensive. In this paper, to detect hate text from roman urdu and urdu text word n-gram method has proposed. Annotated these data and divided into offensive or non-offensive.

Third, extracted the features using character-n gram, forth by applying seventeen classifiers with seven ML techniques it classifies comments into offensive or non-offensive(Akhter et al., 2020). Fifth compare performance of all classifiers and evaluate performance. We used Bayesian model, nearest neighbor, Tree, Random, Regression, SVM, and rule based models. Experiments were conducted and shows model of regression using character n-gram having better performance other than six models. Logiboost shows better work using character trigram on roman urdu obtained 99.2% of F-measure. Simple logistic performs better and obtained 95.8% F-measure on Urdu dataset. In future work, has to apply NN models, FCNN, CNN for detecting offensive in urdu and roman urdu(Akhter et al., 2020).

2.5 Deep learning based CNN model(Janardhana et al., 2021)

D.R. Janardhan et al. designed a method to focus the problem of detecting abusive text and block the user immediately. This paper primarily focuses on automatic detection of abusive comments by using deep learning model(Janardhana et al., 2021). The proposed model has various stages in process of classification. First, the system admin can login and authenticate by giving valid email and password. Second updating the training data, adding new labeled comments to classifier. This can be done by admin. The user can login by using registered email and password. After successful authentication user can post the comments on page. The user generated text to be preprocessed and convert into meaning full tokens(Janardhana et al., 2021). Extracting the features and fed in to the classifier. The classifier can able to classify the comments abusive or not by considering the accuracy. If the comments abusive it will not shows on screen and immediately blocked. Otherwise it will be displayed on screen. Twitter data set has taken and trained the data by using model. In this paper proposed deep learning model CNN classifier achieved 89% of accuracy and it classifies the comments positive, negative and abusive.

Parametric assessment framework

This section depicts comparison of existing methods. Table-1 shows comparison of models based on some parameters as follows. Input language is used to collect the data in the form of native languages. Word weighting scheme is used to calculate the weight of words in the form of unigram, bigram, and trigram etc. classified cluster parameter shows the text has been categorized in to which domain. Data set medium shows that training and tested data collected from which social network. Classifier used to classify the text is abusive and clean. Various types of features selected to identify the offensive content. Result is evaluated by calculating of precision, recall, F-measure and Accuracy

Method	**S hahnoor C. Eshan et al.** (Eshan & Hasan, 2018)	**Discriminative Multinomial Naive Bayes algorithm.**(Tuarob & Mitrpanont, 2017)	**Devanagari Hindi Offensive Tweets (DHOT) Corpus**(Jha et al., 2020)	**Mohammad Pervez Akhter et al**(Akhter et al., 2020)	**Deep learning based CNN model**(Janardhana et al., 2021)
Input Language	Bengali	Thai	Hindi written in Roman	Urdu and Roman urdu	English
Word weighting scheme	TF-IDF	TF,IDF AND TF-IDF	word2vec	character n-gram	Tokenization
Classified cluster	Abusive and non-abusive	Abusive(Rude,figurative,offensive,dirty) or not	hate speech, abusive language, or humorous and non-abusive	offensive and non-offensive	abusive and non-abusive
Dataset medium	Facebook group	Facebook	Twitter	YouTube	Twitter
Classifier	SVM with linear kernel	Naïve bayes	Fast Text	Regression based model	CNN model
Features	unigram, bigram, and trigram	IDF features	character n-gram	character level and word level	Tensor flow
Evaluation	Good	best	Satisfactory	good	good

Table 1: Parametric assessment framework of abusive text summarization approaches

Performance Analysis:

This section depicts measuring performance of classifiers. Table 2 shows the comparison of models based on some parameters as follows. Language is used to collect the text in the form of native language. Feature extraction method is used to calculate the frequency of words in the form of unigram, bigram, and trigram etc. Classifier used to classify the text is abusive and clean. . Result is evaluated by calculating F-measure and Accuracy.

Reference	**language**	**Feature extraction method**	**classification model**	**Accuracy**	**F-measure**
Eshan & Hasan, 2018	Bengali	TfidfVectorizer	SVM linear kernel	90%	
Tuarob & Mitrpanont , 2017	Thai	Idft	DMNB		86%
Jha et al., 2020	Hindi	Word2vec	Fast text	92.20%	
Akhter et al., 2020	Urdu	character n-gram	Simple logistic		95.90%
Janardhana et al., 2021	English		CNN	89%	

Table 2: Performance assessment of classifiers on different languages

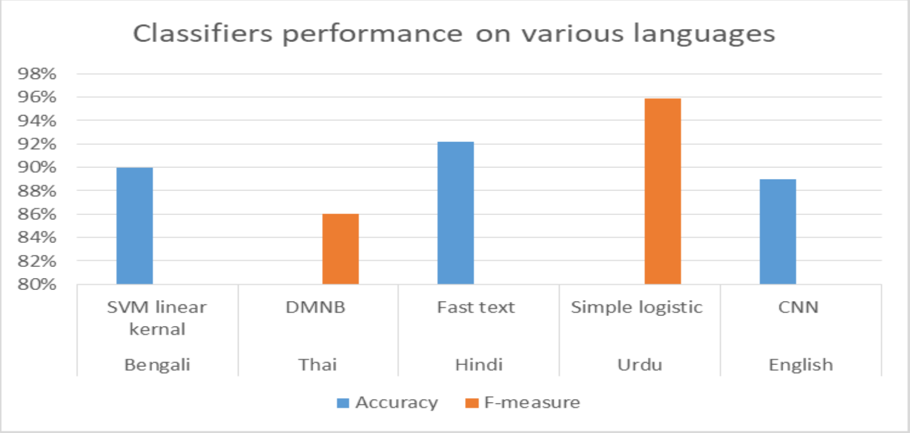

Fig.1 Performance of classifiers in terms of measuring Accuracy, F-measure

Conclusion and Future work

This paper investigates various abusive text detection techniques in different languages using Machine learning and deep learning models. Detecting abusive text in social networks is security to the people. Tfidf, vectorizer features with SVM linear kernel are more effective when compared to count vectorizer features also, MNB performs better in terms of accuracy with bigram and trigram features(Eshan & Hasan, 2018). Discriminative multinomial naïve Bayes algorithm with IDF feature weighted schemes performs best achieved F-measure of 86%.(Tuarob & Mitrpanont, 2017) The fast text model identifying and classifying offensive text from non-offensive which written in devanagari hindi offensive tweet corpus(Jha et al., 2020). This model achieves 92.2% accuracy in identifying abusive comments. By applying seventeen Machine learning classifiers with seven ML models to identify abusive text from urdu and roman urdu text comments and shows regression based model with character n-gram gives better performance in urdu language(Akhter et al., 2020). Logiboost and simple logistic performs better and achieved 99.2% and 95.9% F-measure on both urdu and roman urdu datasets. The figure1 compares Different classifiers on various languages by measuring parameters of Accuracy, F-measure Finally Regression based model of Simple logistic classifier achieving 95.9% of accuracy and detects abusive comments in urdu language.

Future work may lot of research has to be conduct on south languages because it is very challenging to the researchers to identify the abusive content in these languages. More datasets are require to train the model also various syntactic feature methods needed to extract features from different languages. For better results can implement deep learning models CNN, RNN on huge datasets.

References:

Akhter, M. P., Jiangbin, Z., Naqvi, I. R., Abdelmajeed, M., & Sadiq, M. T. (2020). Automatic Detection of Offensive Language for Urdu and Roman Urdu. *IEEE Access*, *8*, 91213–91226. https://doi.org/10.1109/ACCESS.2020.2994950

Alotaibi, A., & Abul Hasanat, M. H. (2020). Racism Detection in Twitter Using Deep Learning and Text Mining Techniques for the Arabic Language. *Proceedings - 2020 1st International Conference of Smart Systems and Emerging Technologies, SMART-TECH 2020*, 161–164. https://doi.org/10.1109/SMART-TECH49988.2020.00047

Eshan, S. C., & Hasan, M. S. (2018). An application of machine learning to detect abusive Bengali text. *20th International Conference of Computer and Information Technology, ICCIT 2017*, *2018-Janua*, 1–6. https://doi.org/10.1109/ICCITECHN.2017.828

1787

Guellil, I., Adeel, A., Azouaou, F., Chennoufi, S., Maafi, H., & Hamitouche, T. (2020). Detecting hate speech against politicians in Arabic community on social media. *International Journal of Web Information Systems*, *16*(3), 295–313. https://doi.org/10.1108/IJWIS-08-2019-0036

Janardhana, D. R., Shetty, A. B., Hegde, M. N., Kanchan, J., & Hegde, A. (2021). Abusive Comments Classification in Social Media Using Neural Networks. *Advances in Intelligent Systems and Computing*, *1165*(May), 439–444. https://doi.org/10.1007/978-981-15-5113-0_33

Jha, V. K., Hrudya, P., Vinu, N. V., Vijayan, V., & Prabaharan, P. (2020). DHOT-Repository and Classification of Offensive Tweets in the Hindi Language. *Procedia Computer Science*, *171*, 2324–2333. https://doi.org/10.1016/j.procs.2020.04.252

Khanam, M. H., Khudhus, M. A., & Babu, M. S. P. (2016). Named Entity Recognition using Machine learning techniques for Telugu language. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, *0*, 940–944. https://doi.org/10.1109/ICSESS.2016.7883220

Mathur, P., Sawhney, R., Ayyar, M., & Shah, R. (2019). *Did you offend me? Classification of Offensive Tweets in Hinglish Language*. 138–148. https://doi.org/10.18653/v1/w18-5118

Melton, J., Bagavathi, A., & Krishnan, S. (2020). DeL-haTE: A Deep Learning Tunable Ensemble for Hate Speech Detection. *Proceedings - 19th IEEE International Conference on Machine Learning and Applications, ICMLA 2020*, 1015–1022. https://doi.org/10.1109/ICMLA51294.2020.00165

Mulki, H., Haddad, H., Bechikh Ali, C., & Alshabani, H. (2019). *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*. 111–118. https://doi.org/10.18653/v1/w19-3512

Naidu, R., Bharti, S. K., Babu, K. S., & Mohapatra, R. K. (2018). Sentiment analysis using Telugu SentiWordNet. *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking, WiSPNET 2017*, *2018-Janua*, 666–670. https://doi.org/10.1109/WiSPNET.2017.8299844

Pratibha, G., & Maneesh, D. (2018). *Parsing Sentiment in Telugu Language*. 99–102.

Priyadharshini, G. (2020). *Detection of Hate Speech using Text Mining and Natural Language Processing*. *9*(11), 2018–2021.

Ranjan, P., Raja, B., Priyadharshini, R., & Balabantaray, R. C. (2016). A comparative study on code-mixed data of Indian social media vs formal text. *Proceedings of the 2016 2nd International Conference on Contemporary Computing and Informatics, IC3I 2016*, 608–611. https://doi.org/10.1109/IC3I.2016.7918035

Sandaruwan, H. M. S. T., Lorensuhewa, S. A. S., & Kalyani, M. A. L. (2019). Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning. *19th International Conference on Advances in ICT for Emerging Regions, ICTer 2019 - Proceedings*, 1–8. https://doi.org/10.1109/ICTer48817.2019.9023655

Sarna, G., & Bhatia, M. P. S. (2017). Content based approach to find the credibility of user in social networks: an application of cyberbullying. *International Journal of Machine Learning and Cybernetics*, *8*(2), 677–689. https://doi.org/10.1007/s13042-015-0463-1

Sreelakshmi, K., Premjith, B., & Soman, K. P. (2020). Detection of Hate Speech Text in Hindi-English Code-mixed Data. *Procedia Computer Science*, *171*(2019), 737–744. https://doi.org/10.1016/j.procs.2020.04.080

Tuarob, S., & Mitrpanont, J. L. (2017). Automatic discovery of abusive thai language usages in social networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10647 LNCS*, 267–278. https://doi.org/10.1007/978-3-319-70232-2_23

Yadav, S. H., & Manwatkar, P. M. (2015). An approach for offensive text detection and prevention in Social Networks. *ICIIECS 2015 - 2015 IEEE International Conference on Innovations in Information, Embedded and Communication Systems*, 3–6. https://doi.org/10.1109/ICIIECS.2015.7193018