Multilingual Cyberbullying Detection System

Rohit Pawar, Rajeev R. Raje
Department of Computer and Information Science
Indiana University-Purdue University Indianapolis, Indianapolis, IN - 46202
{rspawar, rraje}@iupui.edu

***Abstract*— As the use of social media has evolved in recent times, so has the ability to cyberbully victims using it. The last decade has witnessed a surge of cyberbullying – this bullying is not only limited to English but also happens in other languages. A large number of mobile device users are in Asian countries such as India. Such a large audience is a fertile ground for cyberbullies – hence, it is very important to detect cyberbullying in multiple languages. Most of the current approaches to identify cyberbullying are focused on English text, and a very few approaches are venturing into other languages. This paper proposes a Multilingual Cyberbullying Detection System for detection of cyberbullying in two Indian languages – Hindi and Marathi. We have developed a prototype that operates across data sets created for these two languages. Using this prototype, we have carried out experiments to detect cyberbullying in these two languages. The results of our experiments show an accuracy up-to 97% and F1-score up-to 96% on many datasets for both the languages.**

Keywords—Cyberbullying, Machine Learning, Multilingual Cyberbullying for Indian Languages.

I. INTRODUCTION

A. Cyberbullying

Cyberbullying is bullying that takes place in the digital world and can occur on various forums where people view, participate in, or share content. Bullying or harassment can be identified as a repeated behavior and an intent to harm others [1]. Examples of cyberbullying include derogatory, threatening or harassing messages, pictures, audios and videos. Once such content is posted, they live perpetually in the cyber world. Due to the ease of posting such content, cyberbullying empowers a bully to humiliate and hurt the victim in online communities without ever getting recognized. Furthermore, the fear of getting punished or being a social pariah stops victims and bystanders from reporting incidents. Bullying is most common among kids and youngsters. The effects of cyberbullying are often devastating on such population and the result in victims having lower self-esteem. Bullying can also cause many negative effects such as impacts on the mental and physical health [2], depression and anxiety [3], and can lead to suicidal tendencies [4]. As a consequence of such cyberbullying behavior, the victims may miss or even drop out of school. Hence, cyberbullying an epidemic that needs to be controlled quickly and effectively.

B. Countermeasures by Social Media

Social media sites, such as Facebook and Twitter, provide tools and techniques which can help people to report bullying and thus, provide a safe online experience. These include settings to decide the target audience, blocking certain users, and reporting unacceptable behavior and deleting those users. These techniques, although very important, are reactive in nature – i.e., they happen after such someone has already been victimized. By the time, a person reports the content and a corrective action is taken by the authority, many users may read the offensive post; thus, the negative effects (mentioned earlier) may have taken place. Hence, we need an automatic approach that detects cyberbullying behavior promptly and efficiently.

Most of the prevalent approaches to automatically detect cyberbullying (indicated in the next section) focus on English text and associated forums. However, a large number of mobile device users are in Asian countries such as India, China, Japan, and South Korea [5]. For example, in India, there are 1.16 billion mobile device users [5] and they are very active various social media forums such as WhatsApp and use the Indian languages and their features associated with such apps. This sheer volume necessitates the creation of an automatic cyberbullying detection system in other languages.

This paper describes a Multilingual Cyberbullying Detection System for detection of cyberbullying behavior in two Indian languages – Hindi and Marathi. These two languages have 293 million (4.46% of world's population) and 73 million (1.1% of world's population) native speakers [6]. Hence, the proposed system has a potential of creating a significant impact in making online forums safer for the users of these two languages. Hindi and Marathi languages use 'Devanagari' script and hence, some of the words are common in both the language [7]. However, the grammar of both the languages is a bit different.

C. Objectives of the System

Specific objectives for this research are:

- To detect cyberbullying which uses machine learning algorithms to detect bullying messages for English, Hindi and Marathi.
- To examine various machine learning techniques and their effects on the accuracy of detection of cyberbullying messages by empirical evaluations.

II. RELATED WORK

In [8], Haider et al. describe a survey on multilingual cyberbullying detection. They found out that most of the work in this domain is done in English and they attempted cyberbullying detection in Arabic language [9]. In their effort, they used ML learning approach to detect cyberbullying. Their dataset contained 32K tweets; out of which 1800 tweets were bullying ones. They used Support Vector Machine (SVM) and Naïve Bayes algorithms to detect cyberbullying and achieved F1 scores of 92% and 90% respectively.

Ting et al. [10] gathered a dataset from 4 popular social sites in Taiwan. They used Social Network Mining technique to detect cyberbullying. They identified three features from the data: Keywords, Social Network Analysis, and Sentiment. They indicated that sentiment is the most important feature to detect cyberbullying as it helps to

This is the author's manuscript of the article published in final edited form as:

Pawar, R., & Raje, R. R. (2019). Multilingual Cyberbullying Detection System. 2019 IEEE International Conference on Electro Information Technology (EIT), 040–044. https://doi.org/10.1109/EIT.2019.8833846

understand the sentiment/intent of user when he posts message on social media. They used precision and recall as performance measurements. The evaluation results show the precision to be around 79% and the recall around 71%.

In [11], Silva et al. developed a mobile app called 'BullyBlocker'. The main aim of their work was to develop a mobile app on the top of a machine learning model. This app not only helps in cyberbullying detection but also send bullying detection alerts to parents. This app crawls the Facebook feed and messages using the Facebooks API and holds the record of bullying behavior for last 60 days.

In [12], Özel et al. prepared a dataset from Instagram and Twitter messages written in Turkish and then applied machine learning techniques SVM, decision tree (C4.5), Naïve Bayes Multinomial, and k-Nearest Neighbors (kNN) classifiers to detect cyberbullying. They applied information gain and chi-square feature selection methods to improve the accuracy of classifiers. They observed that when both words and emoticons in the text messages are considered as features, cyberbully detection improves. Among the classifiers, Naïve Bayes Multinomial was the most successful one in terms both classification accuracy and running time and they achieved 84% accuracy using it.

In [13], Chen et al. proposed a method called Lexical Syntactic Feature (LSF) for the detection of cyberbullying. For message-level offensive detection, this method heavily relies on BoW (Bag of Words), and the N-Gram techniques. They achieved precision of 98.24% and recall of 94.34% in sentence offensive detection for English.

In our previous work [14], we have described a system which not only detects the cyberbullying in English but also provides distributed infrastructure which is scalable and fault tolerant.

As mentioned earlier, the most of the past work focuses on English and a few other languages (indicated above) but, there is not even single attempt to detect cyberbullying for Indian languages (such as Hindi and Marathi) and that is the aim of this paper.

III. MACHINE LEARNING TECHIQUES

A. An Overview

As mentioned in previous section, Machine Learning (ML)-based classification models are used for detecting cyberbullying. ML is mainly classified into three categories: i) Supervised Learning: in this approach, the mathematical model is built based on data which contains both set of inputs and desired outputs [15]; ii) Unsupervised Learning: in this approach, the model takes set of data as input, and try to find out structure (e.g., grouping or clustering of the data) [15]; and iii) Reinforcement Learning: this approach is concerned with taking suitable actions so as to maximize the reward in particular situation [15].

B. Performance Metrics

Following are the typical performance metrics that are used to evaluate and compare performance of various classifications techniques [16]. In this work, we have used these four metrics to assess the performance of our system:

- Accuracy: This metric measures the number of tweets correctly classified. It is calculated as:

$$\text{Accuracy} = (TP + TN) / T$$

- Precision: This metric measures the number of tweets classified by the algorithm as bullying and actually proved to be bullying tweets. It is calculated as:

$$\text{Precision} = TP / (TP + FP)$$

- Recall: This metric measure how many bullying tweets, out of all available ones, are actually detected by the algorithm. It is calculated as:

$$\text{Recall} = TP / (TP+FN)$$

- F1-Score: This metric computed using the harmonic mean of precision and recall. F1-Score is calculated by the following formula:

$$F = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

where:

TP: Number of True Positive
TN: Number of True Negavtive
FP: Number of False Positive
FN: Number of False Negavtive
T: Total number of tweets

IV. PROPOSED SYSTEM

As already mentioned, the main aim of this paper is to detect cyberbullying behavior in Hindi and Marathi texts appearing on different online forums. Our system employs principles of ML, and thus, as a first step, we had to create a dataset for training and testing the ML models. We created out ML model using python's ML framework i.e., scikit-learn [17].

A. Data Set

1) Data Gathering and Labelling: To train the ML model, we had to collect data from different sources. Data gathering was a challenge since Hindi and Marathi languages have limited resources publically available. Hence, we had to write a scrapper and use APIs to gather the data set for this study. We gathered this data set from multiple sources which include tweets, newspaper reviews, and tourist reviews.

For Hindi language-related study, we obtained data from different domains and topics. These includes movie reviews [18], tour reviews [19], and newspaper reviews [20] on controversial topics such as harassment. The movie review [18] dataset contains 245 reviews; the tour review [19] dataset contains 192 reviews, and we manually collected 184 new paper reviews from [20] on harassment and we tagged those reviews manually. Hence, for the Hindi-related experiments, we gathered and used 621 reviews.

For Marathi language-related experiments, again, we obtained data from different sources and different domains. These included tour reviews [21], newspaper reviews [22] and tweets [23] from the Maharashtra state (state whose official language is Marathi). The Marathi tour review dataset [21] has 106 records; we also collected newspaper reviews from multiple sources [22] containing 196 reviews. Apart from these two sources, we downloaded 508 tweets using the Twitter API [23]. Hence, in all, for the Marathi study, we collected 810 reviews.

Due to context sensitivity of Indian languages, and to ensure correct labelling of sarcastic messages, we manually labelled the messages in both the Hindi and Marathi datasets. We introduced a new attribute called "bullying" (i.e., output label) – if the value of this attribute is "yes", it indicates that

the message is bullying in nature and a value of "no" indicates the non-bullying behavior. This attribute is needed along-with message to train the ML model.

2) Data Pre-processing: Since, we obtained data from multiple sources; it contains lot of unnecessary characters (such as #, @ etc.), stop-words, URLs, punctuations and user ids. So, the first task after data gathering is to remove such unwanted words/characters. For example, here is an instance of the Marathi tweet obtained from Twitter:

"@rohitpawar007489 येणारा काळ सुख समाधानाचा जावो ही सदिच्छा!!! 💐"

This tweet contains a user handle (@rohitpawar007489), an emoji (💐), and a punctuation symbol (!). These entities are unnecessary and not required for training the ML model. This tweet after removing such entities, results into:

"येणारा काळ सुख समाधानाचा जावो ही सदिच्छा".

3) Synthetic Data Generation: After manual tagging of the data set, we realized that dataset of both the language contained approximately 9% of bullying messages. Hence, in order to avoid the data imbalance issue [24], we decided to generate additional instances of bullying messages from the existing instances. To generate the new synthesized data sets, we performed the following steps:

- We stored the pre-processed cyberbullying messages into a list.
- Decide the number of additional instances to be incorporated into the datasets. We decided to double these instances so that resulting dataset will have at least 20% bullying messages.

B. Bag Of Words

We converted the pre-processed string data into the Bag of Words (BoW) format. The BoW format disregards grammar and the order of the words but retains the frequency of the words. The BoW technique is the most common and effective approach used in the text classification problem [25]. We have used the BoW format for all of our experiments. We have also performed 10-fold cross validation for all our experiments. This means that each data point appears only once in the test dataset and 9 times in the training dataset. The purpose of 10-fold validation is, to generalize the model by computing the average error across the folds, no matter how the data is divided [26].

C. System Training and Testing

We have chosen three models, Multinomial Naive Bayes (MNB), Logistics Regression (LR), and Stochastics Gradient Descent (SGD). These algorithms were selected as they perform well on Topic Modeling and Text Classification, as indicated in our past work [14] as well as in literature (please refer to Section II). These machine-learning algorithms were trained to create models that were used for the classification of the cyberbullying tweets. We used 80% of the data for the training purpose and remaining 20% for the testing purpose. Results of our experiments are discussed in next section.

V. RESULTS

As already mentioned (in Section Data Set), we have obtained data from multiple sources and conducted all the experiments using the Multinomial Naive Bayes (MNB), Logistics Regression (LR), and Stochastics Gradient Descent (SGD) algorithms. In addition, we carried out experiments with and without the synthesized data set. Results obtained for both the Hindi and Marathi language-related datasets are discussed in next two paragraphs. In past, we have carried out ML experiments for English text [14]. We have obtained those results as well from [14] for the comparison purpose.

Results of our experiments for Hindi text are as shown in Table I, for Marathi texts are shown in Table II, and for English texts are shown in Table III [14]. These results indicate that Logistics Regression (LR) outperforms SGD and MNB in all the languages. In addition, performance of all the ML algorithms is improved by generating additional data using data synthesizing technique. MNB has the assumption that every feature is independent but that is not possible in real situations [27] – thus, it does not outperform LR in our experiments as well. As reported in [28], LR performs well for the binary classification problem and continues to work better as data size grows [28]. LR updates a set of parameters in an iterative manner and tries to minimize the error function whereas, SGD uses one sample and uses the close approximation to update the parameters. Hence, SGD performs faster but error is not as minimized as in LR [29]. So, it is not surprising that LR outperforms the other two approaches in our experiments as well.

Results in Tables I, II and III also show that our model performs as expected, even when we add more data to create the synthesized dataset. Accuracy is not good measure to compare performance of the model especially when dataset is imbalanced, hence, we use F1-score as the performance measure. Results, with the synthesized dataset (Tables I, II, and III), show that the addition of more data to our dataset improves the F1 score. This indicates that our model is generalized and performs better on both the classes (i.e., bullying and non-bullying) than when it is created with the imbalanced (i.e., actual) dataset.

VI. CONCLUSION AND FUTURE WORK

This paper has provided a multilingual cyberbullying detection approach for detecting cyberbullying in messages, tweets and newspaper review for two Indian languages. Results of our experiments shows that Logistics Regression outperforms all other algorithms on these datasets. Also, generating synthesized data could help us improve performance of our system. Results of our study show that our systems perform well across two languages and different domains and hence, it can be used to detect cyberbullying for other Indian languages as well.

Many future extensions of our works are possible. These are as follows:

- We would like to validate this approach on very large datasets.
- We would like to provide language inputs and detect sentiment, and sarcasm associated with it.
- Explore other approaches such as Natural Language Process (NLP) and, using translator and compare performance of different approaches.
- Integrate our approach into the distributed prototype created in our previous work [14] to achieve collaborative cyberbullying detection in real-time.

TABEL I. RESULTS FOR HINDI DATASET

No.	Dataset	Algorithm	Synthesize Data	Accuracy	Precision	Recall	F1-Score
1	Movie Reviews	SGD	No	0.7346	0.7502	0.7347	0.7347
		MNB	No	0.6734	0.6735	0.6735	0.6735
		LR	No	0.7346	0.7346	0.7346	0.6933
		SGD	Yes	0.7391	0.7801	0.7391	0.7441
		MNB	Yes	0.7681	0.7636	0.7681	0.7631
		LR	Yes	0.7826	0.7826	0.7826	0.7626
2	Tour Reviews	SGD	No	0.7948	0.7985	0.7949	0.7946
		MNB	No	0.8717	0.8729	0.8718	0.8718
		LR	No	0.7179	0.7123	0.7187	0.7134
		SGD	Yes	0.9322	0.9322	0.9322	0.9322
		MNB	Yes	0.9491	0.9527	0.9492	0.9479
		LR	Yes	0.9452	0.9435	0.9425	0.9415
3	Newspaper Reviews	SGD	No	0.4594	0.4669	0.4595	0.4618
		MNB	No	0.3513	0.3563	0.3585	0.3523
		LR	No	0.5135	0.5285	0.5135	0.5149
		SGD	Yes	0.7719	0.7770	0.7719	0.7742
		MNB	Yes	0.8070	0.8050	0.8060	0.7970
		LR	Yes	0.9122	0.9126	0.9123	0.9089

TABEL II. RESULTS FOR MARATHI DATASET

No.	Dataset	Algorithm	Synthesize Data	Accuracy	Precision	Recall	F1-Score
1	Tour Reviews	SGD	No	0.9523	0.9549	0.9524	0.9483
		MNB	No	0.9523	0.9643	0.9524	0.9551
		LR	No	0.8571	0.8524	0.8563	0.8588
		SGD	Yes	0.9024	0.9219	0.9024	0.9092
		MNB	Yes	0.9512	0.9652	0.9512	0.9546
		LR	Yes	0.9756	0.9235	0.9574	0.9575
2	Twitter Tweets	SGD	No	0.8157	0.9433	0.8158	0.8749
		MNB	No	0.7894	0.9423	0.7895	0.8591
		LR	No	0.8236	0.9323	0.8236	0.8954
		SGD	Yes	0.9482	0.9536	0.9483	0.9484
		MNB	Yes	0.9655	0.9680	0.9655	0.9656
		LR	Yes	0.9655	0.9648	0.9668	0.9688
3	Newspaper Reviews	SGD	No	0.7037	0.7037	0.7037	0.7037
		MNB	No	0.7777	0.7156	0.7778	0.7454
		LR	No	0.8518	0.8186	0.8578	0.8234
		SGD	Yes	0.9148	0.9172	0.9149	0.9143
		MNB	Yes	0.9361	0.9367	0.9362	0.9360
		LR	Yes	0.9574	0.9598	0.9527	0.9572

TABLE III. RESULTS FOR ENGLISH DATASET [14]

No.	Algorithm	Synthesize Data	Accuracy	Precision	Recall	F1-Score
1	MNB	No	0.8780	0.8865	0.8780	0.8792
		Yes	0.8845	0.8974	0.8895	0.8845
2	SGD	No	0.9232	0.9257	0.9045	0.9177
		Yes	0.9352	0.9365	0.9135	0.9263
3	LR	No	0.9311	0.9311	0.9312	0.9307
		Yes	0.9424	0.9421	0.9438	0.9412

REFERENCES

[1] S. Hinduja and J. Patchin, "Cyberbullying: Neither an epidemic nor a rarity," European Journal of Developmental Psychology, vol. 9, no. 5, pp. 539–543, 2012.

[2] K. Kumpulainen, E. R¨as¨anen, and K. Puura, "Psychiatric disorders and the use of mental health services among children involved in bullying,"Aggressive behavior, vol. 27, no. 2, pp. 102–110, 2001.

[3] W. Craig, "The relationship among bullying, victimization, depression, anxiety, and aggression in elementary school children," Personality and individual differences, vol. 24, no. 1, pp. 123–130, 1998.

[4] A. Klomek, A. Sourander, and M. Gould, "The association of suicide and bullying in childhood to young adulthood: a review of cross-sectional and longitudinal research findings", The Canadian Journal of Psychiatry, vol. 55, no. 5, pp. 282–288, 2010.

[5] Mobile phone users in countries; Wiki. Retrieved from https://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use, 2/16/19.

[6] Language Native Speakers Wiki. Retrieved from https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers, 2/16/19.

[7] Difference between Hindi and Marathi. Retrieved from https://www.quora.com/What-is-the-difference-between-Marathi-and-Hindi, 4/19/19.

[8] B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying detection: A survey on multilingual techniques," in European Modelling Symposium (EMS), pp. 165–171, Nov 2016.

[9] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbul-lying detection system: Detecting cyberbullying in arabic content," in 2017 1st Cyber Security in Networking Conference (CSNet), pp. 1–8, Oct 2017.

[10] I. Ting, W. Liou, D. Liberona, S. Wang, and G. Bermudez,"Towards the detection of cyberbullying based on social network mining techniques," in International Conference on Behavioral, Economic,Socio-cultural Computing (BESC), pp. 1–2, Oct 2017.

[11] Y. Silva, C. Rich, and D. Hall, "Bullyblocker: Towards the identification of cyberbullying in social networking sites," in IEEE/ACM International Conference on Advances in Social Networks Analysis andMining (ASONAM), pp. 1377–1379, Aug 2016.

[12] S. Zel, E. Sara, S. Akdemir, and H. Aksu, "Detection of cyberbullyingon social media messages in turkish," in International Conference on Computer Science and Engineering, pp. 366–370, Oct 2017.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language insocial media to protect adolescent online safety," in International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing, pp. 71–80, Sep. 2012.

[14] R. Pawar, Y. Agrawal, A. Joshi, R. Gorrepati, and R. Raje, "Cyberbullying detection system with multiple server configurations," in 18th IEEE International Conference on Electro/Information Technology, pp. 0090–0095, May 2018.

[15] Machine Learning Basics. Retrieved from https://www.edureka.co/blog/what-is-machine-learning/, 2/16/19.

[16] Twitter sentiment algorithms benchmarking precision, recall, f-measures. Retrieved from https://www.linkedin.com/pulse/20141126005504-34768479-twitter-sentiment-algos-benchmarking-precision-recall-f-measures, 2/16/19.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn:Machine learning in Python,"Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[18] Hindi movie reviews, Retrieved from: http://www.cfilt.iitb.ac.in/resources/senti/download.php?get=hplc.zip, 2/16/19.

[19] Hindi tour reviews, Retrieved from: http://www.cfilt.iitb.ac.in/resources/senti/download.php?get=hplctour.zip, 2/16/19.

[20] Hindi news paper reviews, Retrieved many pages from the https://navbharattimes.indiatimes.com/ website, Jan-Feb 2019.

[21] Marathi tour reviews, Retrieved from http://www.cfilt.iitb.ac.in/resources/senti/download.php?get=mplctour.zip, 2/16/19.

[22] Marathi news paper reviews, Retrieved many pages from the https://maharashtratimes.indiatimes.com/ website, Jan-Feb 2019.

[23] Twitter APIs, Retrived from https://developer.twitter.com/en/docs.html, 2/16/19.

[24] Data imbalance issue and techiques to overcome, Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/, 2/16/19 .

[25] Bag of words model, Retrieved from https://machinelearningmastery.com/gentle-introduction-bag-words-model/, 2/16/19.

[26] R. Kohavi, "A study of cross-validation and bootstrap for accuracyestimation and model selection," in IJCAI, vol. 14, pp. 1137–1145, Montreal,Canada, 1995.

[27] Naïve bayes and Logistic Regression Comparision. Retrieved from http://dataespresso.com/en/2017/10/24/comparison-between-naive-bayes-and-logistic-regression/, 2/16/19.

[28] Why Logistic Regression over Naïve Bayes. Retrieved from https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c, 2/16/19.

[29] Logistic Regression vs Stochastic Gradient Descent. Retrieved from https://www.quora.com/In-scikit-learn-what-is-the-difference-between-SGDClassifer-with-log-loss-and-logistic-regression, 2/16/19.