2018 2nd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET)
27-28 October 2018, Chittagong, Bangladesh

Detecting Abusive Comments in Discussion Threads Using Naïve Bayes

Md. Abdul Awal
Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
awal.kuet@yahoo.com

Md. Shamimur Rahman
Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
shamimur052@gmail.com

Jakaria Rabbi
Department of Computer Science and Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
jakaria.rabbi@yahoo.com

***Abstract*—Comments are supported by various websites and provide a simple approach to increment user involvement. Users can generally comment on different types of media such as: social networks, blogs, forums and news articles. As discussions increasingly move toward online forums, the issue of insulting and abusive comments is becoming prevalent. In addition,a lots of comments are available due to these social media. Hence, it is not feasible for a human moderator to check each comments one by one and flag them as abusive or not abusive. For this reason, an automated classifier which is quick and efficient is necessary to detect such type of comments. To fulfill above purpose, in this paper a Naïve Bayes classifier is designed to detect abusive comments expressed in Bangla. Using a training corpus collected from "Youtube.com", the Naïve Bayes classifier is employed to categorize comments as abusive or not abusive. Finally, the performance is evaluated by using 10-fold cross-validation on unprocessed data.**

Keywords—Abusive comments, Naïve Bayes, machine learning, text classification, 10-fold cross-validation.

I. INTRODUCTION

Human interaction with social networks, blogs, forums and online news portals has been increased drastically in the previous couple of years. Social networks, blogs, forums and online news portals unite users to form a strong association generally based on a way of communication via messages, chats and comments. Comments capacitate a casual and interactive way of providing personal point of view. Generally commenters are unopposed to express their sentiments, share their responses, and offer their learning. Readers obtain additional facts over the article from comments and usually they also react to comments by giving reply. Users generally utilize "thumbs up" or "thumbs down" sign in order to short response to a comment [15]. In addition detailed responses are also feasible – prompting to "comment threads". Consequently comments give a feeling of group interest by a low passage obstruction.

The comments can appear as any composed content whether it is in English, Bangla or something else. Most of the time, the commenting framework is an essential part of making a group in a website. This framework normally allows anonymous posting that gives users the chance to misconduct the commenters or posters on the framework. So simply like some other community feature, comments are defenseless to manhandle. That's why, identification and blocking of abusive comments are indispensable for the transparency of comments. The consequences of abusive comments are multifarious[15].

- Since readers need to filter through comments spam to get good comments, they can lose their enthusiasm to a website.
- Normally commenters are discouraged to comment in an environment which may be full of spam and their comments are probably going to be suffocated in an ocean of spam.
- The owners or proprietors of sites may observe less user involvement and graduallypoor quality traffic.

Abuse or misconduct on a commenting framework varies from spam to comments which are infelicitous. Users often recognize this content highly invective. As a result, the websites can obtain negative feedbacks from users and also lose their traffic. So the moderators have a critical undertaking in securing the fairness of a website [21]. They impose particular rules and regulations about what types of comments are allowed to post. Suppose, an abusive comment could assault a user utilizing pejorative terms, then it is the responsibility of a moderator to decide if this comment should be allowed or not for posting. Generally human being plays the role of a moderator, who have to read each of the comments to categorize them as abusive or not. However, manually reviewing and detecting offensive comments are tiresome and time killing task and hence not feasible, reliable and usable in practical sense.

In order to identify and block abusive and offensive contents of a website, some automated software such as "Appen" and "Internet Security Suite" have been used [2]. These software packages juststopped webpages from loading into a web browser which contained scurrilous contents. The methodboth interrupts the readability and usability of website and fails to identify exquisite insulting contents. The purpose of this research is to detect abusive comments expressed in Bangla. At first, the dataset of English comments is collected from "Youtube.com" [4]. Then the annotated Bangla dataset is generated from this collected dataset. Naïve Bayes classifier is trained on this dataset. Finally 10-fold cross-validation technique is applied to measure the accuracy of the classifier.

The organization of the paper can be described as below: previous works are presented in section II and research methodology is described in section III.Section IV shows the results of the experimental analysis. Finally, the conclusion is presented in section V.

978-1-5386-8524-2/18/$31.00 ©2018 IEEE

II. Existing Works

The task of textual annoyance or abusive commentsidentification in text has been marked by scientists as a classification task. Abusive comments classification research with machine learning began with Yin et al.'s [5] paper. The authors proposed a supervised machine learning technique to detect textual harassment, in which texts are illustratedbased on word frequency features, sentiment features and features whichtake the similarity to neighboring posts. One of the first works to address abusive language was [6], which used a supervised classification technique in conjunction with N-gram, manually developed regular expression patterns, contextual features that take into account the abusiveness of previous sentences.

Dinakar et al. [7]collected dataset from YouTube videos in different topics that contain comments and applied binary and multiclass classifiers. The experimental results indicate that topic-sensitive binary classifiers enhanced the performance of generic multi-class classifiers. Dadvar et al. [8] applied a rule-based expert system, a supervised machine learning model, and a hybrid approach to automatically detect cyberbullying. The author showed that the expert system performs better than the machine learning and hybrid approach model. Nahar et al. [9] proposed a semi-supervised leaning methodwhich will enlarge training data samples and use a fuzzy SVM algorithm. The improved training method automatically extracts and increases training set from the unlabeled streaming text, while learning is conducted by utilizing a little training set provided as an initial input. From the experimental results it is observed that the proposed improvedtechniqueperforms better than all other techniques, and is applicable in the practicalscenarios, whileenough labelled datasetis unavailable for training.

In [10] authors developed and applied a new method to annotate cyberbullying, which indicates the presence and cruelty of cyberbullying, a post author's role (harasser, victim or bystander) and a number of fine-grained categories related to textual harassment such as insults and threats. The experimental results shown the possibility of fine-grained cyberbullying detection.Reynolds et al. [11] collected data from social networking site"FromSpring". They applied machine learning algorithm on this dataset. Amazon Web service Turk was usedto label the collected data. In case of identifying true positives, the accuracy of their technique was 78.5%. Altaf Mahmud et al [12] created a set of semantic rules to distinguish factual and insultingcomments by parsing comments, but they did notconsider direct involvementof participants and nonparticipants.

Razaviet al [13] proposed an automatedabusive content detection approach that extracts features at various conceptual levels using bag-of-words model and applies multilevel classification to detect flame in text. Most recently, Nobatacollected a corpus of Yahoo! Finance and News comments to detect abusive language [14]. The author extracted character N-gram, linguistic, syntactic, and distributional semantic features from this dataset to train the proposed model. Kant et al. [15] developed an abusive content detection framework based on frequent subsequence mining. The authors enhanced the PRISM algorithm to obtain a new algorithm namedmcPRISM that can mine frequent sequences from abusivecontents with expected level of accuracy.Chavan et al. [16] included pronouns, skip-gram, TFeIDF, and N-grams as extra features to improve the accuracy of their proposed model.

A Lexical Syntactic Feature (LSF) based approach was proposed by Chen et al. [2] to detect abusive contents and identify probable offensive users in social media. The authors included a user's writing style, structure and specific textual harassing contents as features to predict the user's probability to send out abusive contents. In case of offensive sentence detectionprecision and recall value were 98.24% and 94.34% respectively andin case of offensive user detection precision and recall value were 77.9% and 77.8% respectively. Xiang et al. [22] applied machine learning and topic modeling approaches to identify profanity-related abusive contents on Twitter. They acquired a true positive rate of approximately 75%, outperforming keyword-based techniques. In [1], the authors proposed an approach to extract opinion from text expressed both in English and Bangla. In this purpose they used Naïve Bayes classifier to extract the opinion. Three levels such as weak, steady and strong were used as the task of opinion mining.

Most of the research works mentioned above only deal with the detection of abusive comments expressed in English. But the purpose of this research is to detect abusive comments expressed in Bangla.

III. Methodology

Offensive language detection in social networks, blogs, forums and news articles is very complicatedjob. The textualsubstances in thiscircumstance are informal, not structured and even incorrectly spelled [2]. While protective techniques accepted by various websites areinadequate, scientists have studied efficient ways to detectinvective contents utilizing text mining techniques.

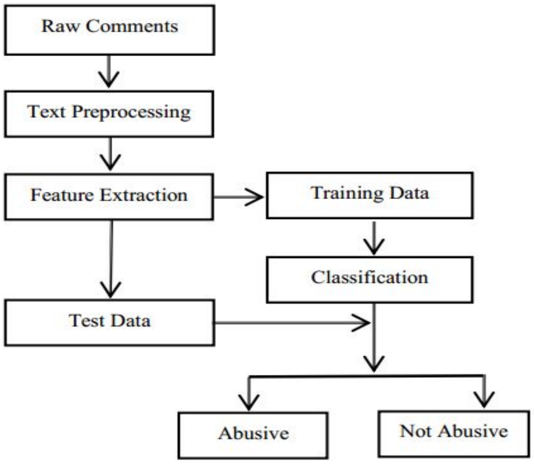

Fig. 1. The flowchart for detecting abusive comments.

To serve this purpose, Naïve Bayes classifier is used in this research to detect abusive comments expressed in Bangla. The methodology to break down information from data needs the majorsteps, which are: 1) data acquisition and preprocess, 2) feature extraction, and 3) model selection. The significant difficulties of utilizing text miningapproach to distinguishhostile contents depend on the feature selection stage. All the steps of the methodology are depicted in Fig. 1.

A. Data acquisition and preprocess:

Abusive contents classification is as yet a comparatively recent research topic in NLP and there are legitimate and privacy issues with making this information public. That's why, few datasets have been curated particularly for this problem. In this study a comment is considered as abusive if either the primarymotive is insult or it retainsinvective or offensive words, phrases or languages. As Bangla dataset for abusive comments detection isunavailable for research, the Bangla dataset is generated from English dataset in two different ways: 1) direct translation to Bangla and 2) dictionary based translation to Bangla [1], [22]. The translation is done by"Google Translator". For this purpose, the English dataset is collected from "Youtube.com" [7], [21]. The dataset contains 2665 instances or English comments. Among them, 1451 were labeled as not abusive or positive comments. The remaining, 1214 comments were marked as abusive or spam. In order to train Naïve Bayes classifier, the dataset must be converted into feature vector. Hence, various natural language processing techniques such as normalization and stemming are applied to remove unwanted strings,like URLs, IP addresses, or other special array of characters [7], [16], [20]. After preprocessing the actual Bangla dataset is generated from this preprocessed English dataset. Table I shows some samples of Bangla comment.

TABLE I. TRANSLATION FROM ENGLISH TOBANGLA.

English	Translated Bangla
sorry but you are just a bitch lady...	দুঃখিত কিন্তু আপনি শুধু একটি দুশ্চরিত্রা মহিলা ...
your head is full of shit	আপনার মাথায় গোবর ভরা
you're a tail-less monkey	তুই একটা লেজকাটা বানর
you are a pig	আপনি একটা শুয়ার

B. Feature extraction:

To train Naïve Bayes classifier, commentsmust be transferred into feature vector. So as toextract features from text, it is necessary to partition text into chunk which is called tokenization. The features are consists of tokens which are collected from text after tokenization. Then every part oftext is reduce to a vector of tokens, where 1 denotes the presence of that token and 0 denotes the absence of that token in a document. The preprocessed dataset is then prepared into a bag-of-word (BOW) vector that calculates the occurrence of a particular word in a particular comment.

C. Model selection:

A classifier can be utilized to moderate a website and operates quicker than having human moderators or users to flag comments. The taskrelated to the dataset applied in this paper is binary classification. For this reason, Naïve Bayes approach is chosen for the classification of abusive comments. The Naïve Bayes approach is easy to implement and computationally efficient. Naïve Bayes is a subset of Bayesian decision theory. The Bayes Theorem enables us to compute the likelihood of an occurrence that prompt to a result. Bayes Theorem states that:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \qquad (1)$$

Suppose, $\mathrm{C_i}$ is the set of i classes and D is a document to classify. The probability of a class C given a document D can be computed by Bayes Theorem as follows:

$$P(C|D) = \frac{P(D|C) \times P(C)}{P(D)} \qquad (2)$$

$$(3)$$

The probability for each class in $\mathrm{C_i}$ is calculated an consider the class that is attached with the highest probability to obtain the classification. The probability of a class P(C) is simply the probability of the class C. If D_C is the total no of documents in class C and D_T is the total no of document in dataset then

$$P(C) = \frac{D_C}{D_T}$$

The ratio of the number of documents in the class C to the aggregate number of whole documents of all classes denotes the probability that a document is in class C. If D is expanded into individual features, then probability of a document P(D|C) can be calculated as follows [26]:

$$P(D|C) = P(d_{0,}d_{1,}d_{2,} \dots , d_{n,}|C) \qquad (4)$$

The assumption is that all the words are independently likely, and somewhat named conditional independence and the probability is calculated as [26]:

$$P(w_0|C_i) * P(w_1|C_i) * P(w_2|C_i) * \dots * P(w_n|C_i) \qquad (5)$$

To calculate the conditional probability for each class, following pseudo code is used[26]:

```
calculate the no of comments in every class
for each training comment
{
   for everyclass
     {
          if (token_value = 1)
             {
                    token_count = token_count + 1
             }
          total_token_count = total_token_count + 1
     }
   for every class
{
          forevery token
             {
                    conditional_probability = token_count /
                    total_token_count
             }}
return conditional_probability
}
```

IV. EXPERIMENTAL ANALYSIS

All the outcomes from the implementation of Naïve Bayes are given in this section. A comment is classified into two polarities: abusive and not abusive. In Naïve Bayes, the probabilistic value is used to determine the class level. Assume that, the conditions for the probability of a bit of informationassociating to class 1 is p1(x, y) and class 2 is p2(x,y). To classify a new bit of information with features (x,y) the following rules are used:

If p1(x, y) > p2(x, y), then the class label is 1.

If p2(x, y) > p1(x, y), then the class label is 2.

From (5) it is noticed that, a bunch of probabilities are multiplied together to find out the probability that a document associates to a particular class. Multiplication of too many small numbers causes underflow or produces an incorrect result. To solve this problem, natural logarithm of each probabilities are multiplied together to get the actual probability.Because it is known that, $\ln(x * y) = \ln(x) + \ln(y)$. It helps us to minimize the effect that comes from underflow or round-off error problem. Fig. 1 [26] plots two functions, $f(x)$ and $ln(f(x))$. From the Fig. 2 it is observed that, both the functions increase and decrease in the same areas and they have their peaks in the same areas.This alsoindicatesthat the natural logarithm of a function can be utilizedasreplacement of a functiontofind the maximum value of that function.

The dataset used in this paper, contains 2665 instances or English comments. Then 10-fold cross-validation technique is applied to evaluate theperformance of the predictive model. The main dataset is divided into a training set to train the model, and a test set to evaluate it. The dataset is arbitrarily divided into 10 equivalent subsets. Among the 10 subsets, a single subset is randomly choose totest the accuracy of the model, and other 9subsets are utilized as training data to train the model. The cross-validation process is reiteratedten times (the folds), with each of the 10 subsets used only once as the validation data. Finally, a single estimation is produced from the average of 10 results.

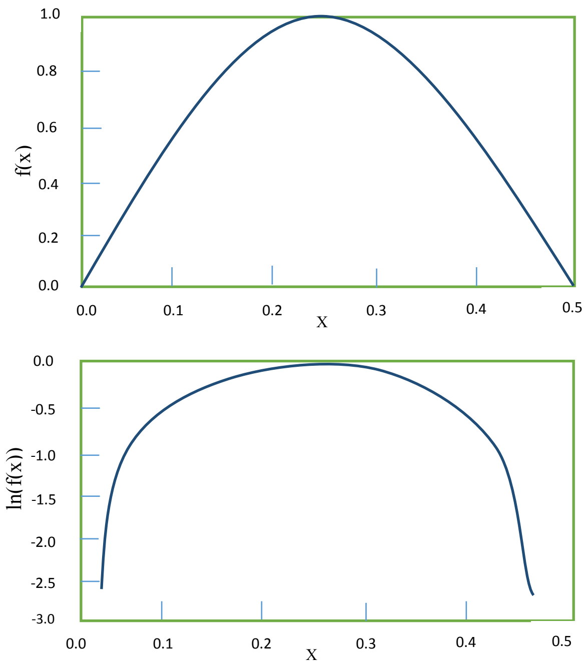

Fig. 2.Arbitrary functions f(x) and ln(f(x)) increasing together.

Precision-Recall metric is used to measure the accuracy of Naïve Bayes classifier. Precision (P) is defined as the ratio of the true positives$(T_p)$to the total number of true positives plus the number of false positives (F_p). That is:

$$P = \frac{T_p}{T_p + F_p} \tag{6}$$

Recall (R) is defined as the ratio of the true positives (T_p) to the total number of true positives plus the number of false negatives(F_n). That is:

$$R = \frac{T_p}{T_p + F_n} \tag{7}$$

Accuracy is measured by the percentage of comments in the test set that the classifier correctly labels. That is:

$$Accuracy(A) = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100 \tag{8}$$

The F_1 score is also used to measure the accuracy of a classifier.It takes into account both the precision P and the recall R of the test to calculate the score.Theweighted average of the precision and recall value is used to define F_1 score.The value 1 and 0 respectively indicates the best case and worst case scenario of F1 score. F_1 score is calculated as follow:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

The confusion matrix of experimental analysis is presented in table II. The experimental results presented in table III show that the precision and the recall value to detect abusive comments in discussion threads are 0.81 and 0.77 respectively. The table also shows that, overall accuracy and F_1score of the Naïve Bayes classifier are 80.57% and 0.39 respectively.

TABLE II. CONFUSION MATRIX OF EXPERIMENTALANALYSIS.

		Actual Class	
Predicted Class		Abusive	Not Abusive
	Abusive	103 True Positives	24 False Positives
	Not Abusive	30 False Negatives	121 True Negatives

TABLE III.RESULT OF EXPERIMENTAL ANALYSIS.

Precision	Recall	Accuracy	F1 Score
$P = \frac{103}{103 + 24} = 0.81$	$R = \frac{103}{103 + 30} = 0.77$	$A = \frac{103 + 121}{103 + 121 + 30 + 24} \times 100 = 80.57\ \%$	$P = \frac{0.81 * 0.77}{0.81 + 0.77} = 0.39$

TABLE IV.EXCEPTIONAL CASE OF ABUSIVE COMMENTS.

English	Bangla	Actual Polarity	Detected Polarity
Those who are prostitutes treated disdainfully in our society	বেশ্যাবৃত্তি যারা করে তাদেরকে আমাদের সমাজে ঘৃণার চোখে দেখে	Not Abusive	Abusive
When talking, it is a very bad thing to be cursed by someone as illegitimate, bastard, prostitute and a son of a bitch.	কথা বলার সময় কাউকে জারজ, হারামজাদা, পতিতা এবং কুত্তার বাচ্চা বলে গালি দেয়া খুব খারাপ কাজ।	Not Abusive	Abusive

There are some cases when Naïve Bayes classifier fails to detect abusive comments. Table IV shows such type of comments among them that cannot be accurately detected by

Naïve Bayes classifier. Here “বেশ্যাবৃত্তি”, “জারজ”, “হারামজাদা”, “পতিতা”, “কুত্তারবাচ্চা”, and “ঘ্ণা” are some extensively used vulgar words in Bangla Language. But in the above comments the overall meaning that is semantic of these words is not offensive. In order to flag such type of comments as not abusive, the semantic of these comments must be considered.

V. Conclusions

As the volume of online user yielded contents are rapidly increasing, it is essential to apply accurate and automated techniques to detect abusive contents. Hence, Naïve Bayes classifier is applied in this paper, to automatically detect abusive comments in discussion threads. To achieve the goal, the dataset is collected from “Youtube.com”. Some pre-processing techniques are applied on this collected dataset to clean unwanted texts and prepare the dataset for training the classifier. As a significant effort, the technique has acquired excellent accuracy to detect abusive comments in discussion threads which is depicted in the experimental analysis section. In case of abusive comments detection, the technique used in this paper will minimize the editorial efforts of a human moderator by an order of magnitude. A future plan is to detect abusive comments including the article it references, any comments preceding or replied to, as well as information about the commenter’s past behavior or comments.

References

[1]. K. M.A. Hasan, M. S. Sabuj, Z. Afrin, “Opinion Mining using Naïve Bayes”, In: IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), pp. 511-514, IEEE.

[2]. Y. Chen, Y. Zhou, S. Zhu, H. Xu,“Detecting offensive language in social media to protect adolescent online safety”, In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom), pages 71–80. IEEE, 2012.

[3]. C. Nobata, J. R. Tetreault, A. Thomas, Y. Mehdad, Y. Chang,“Abusive language detection in online user content”, In Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, pages 145–153, 2016.

[4]. M. M. Nabi, M. T.Altaf, S. Ismail,“Detecting Sentiment from Bangla Text using Machine Learning Technique and Feature Analysis”, International Journal of Computer Applications 153(11):28-34, November 2016.

[5]. D. Yin, Z. Xue, L. Hong, B.D. Davison, A. Kontostathis, L. Edwards,“Detection of harassment on web 2.0”, In: Content Analysis in the WEB 2.0, (CAW2.0) Workshop at WWW, Madrid, Spain, 2009.

[6]. S. O. Sood, J. Antin, E. F. Churchill,“Using crowdsourcing to improve profanity detection”, In AAAI Spring Symposium: Wisdom of the Crowd, 2012.

[7]. K. Dinakar,R. Reichart, H. Lieberman,“Modeling the detection of textual cyberbullying”, Workshop on the Social Mobile Web in 5th International AAAI Conference on Weblogs and Social Media, Spain 2011.

[8]. M. Dadvar, D. Trieschnigg, F. D. Jong, “Experts and machines against bullies: A hybrid approach to detect cyberbullies”, In: Canadian Conference on Artificial Intelligence, Springer (2014) 275–281.

[9]. V. Nahar, S. Al-Maskari,X. Li, C. Pang, “Semi-supervised learning for cyberbullying detection in social networks”, In: ADC, Springer (2014) 160–171.

[10]. C. V.Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. D. Pauw, W. Daelemans,V. Hoste, “Detection and fine-grained classification of cyberbullying events”, In: Recent Advances in NLP Conference (RANLP), (2015) 672–680.

[11]. K. Reynolds, A. Kontostathis, L. Edwards,“Using Machine Learning to Detect Cyberbullying”,10th International Conference on Machine Learning and Applications and Workshops (ICMLA), 2011, vol.2, no.pp.241,244,18-21Dec.2011.

[12]. A. Mahmud,K. Z. Ahmed, M. Khan,“Detecting flames and insults in text”, In: Proceedings of the Sixth International Conference on Natural Language Processing (2008).

[13]. A. H. Razavi, D. Inkpen,S. Uritsky, S. Matwin, “Offensive language detection using multi-level classification”, In: Proceedings of the 23rd Canadian Conference on Artificial Intelligence, pp. 16–27 (2010).

[14]. G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose,“Detecting offensive tweets via topical feature discovery over a large scale twitter corpus”, In Proc. CIKM, pages 1980–1984, New York, NY, USA, 2012. ACM.

[15]. R. Kant, S. Sengamedu, K. Kumar,“Comment spam detection by sequence mining”, In WSDM, pages 183–192. ACM, 2012.

[16]. V. S. Chavan,S. Shylaja, “Machine learning approach for detection of cyberaggressive comments by peers on social media network”, In Advances in computing, communications and informatics (ICACCI), 2015 International Conference on (pp. 2354e2358), IEEE.

[17]. A. Das, S. Bandyopadhyay,“SentiWordNet for Bangla”, February, 23th to 24th, In Knowledge Sharing Event-4: Task 2: Building Electronic Dictionary Mysore, 2010.

[18]. K. M. A. Hasan, S. Islam,G. M. Mashrur-E-Elahi, M. N.Izhar,“Sentiment Recognition from Bangla Text”,DOI: 10.4018/978-1 4666-3970-6.ch014, pp. 1-10, 2010.

[19]. F. K. Ventirozos, I. Varlamis, G.Tsatsaronis, “Detecting aggressive behavior in discussion threads using text mining”, 18th International Conference on Computational Linguistiscs and Intelligent Text Processing, June, 2017.

[20]. S. Chowdhury, W. Chowdhury, “Performing sentiment analysis in Bangla microblog posts”, in: Proceedings of International Conference on Informatics, Electronics & Vision, 2014, pp. 1–6.

[21]. Maus, Adam. “SVM approach to forum and comment moderation.” Class Projects for CS (2009).

[22]. Google Translator: https://translate.google.com.

[23]. https://www.frontgatemedia.com/a-list-of-723-bad-words-to-blacklist-and-how-to-use-facebooks-moderation-tool/

[24]. https://github.com/wooorm/profanities/blob/HEAD/support.md

[25]. http://www.cs.cmu.edu/%7Ebiglou/resources/bad-words.txt

[26]. Peter Harrington, “Machine Learning in Action”, Manning Publications Co.