An Enhanced Ensemble Classifier for Hate and Offensive Content Identification

Rajalakshmi R, Yashwant Reddy B
School of Computing Science and Engineering
Vellore Institute of Technology
Chennai, India
{rajalakshmi.r@vit.ac.in,}{byashwanth.reddy2016@vitstudent.ac.in}

ABSTRACT: *Recent advancements in the Internet technologies have made a tremendous change in the social media. Hate Speech is an attack that is directed towards a group of people based on their religion, gender, colour etc. The offensive content in social media poses a threat to democracy. As these kind of hate speech and offensive content on the web increases day by day, manually monitoring or controlling such hate crimes is a highly challenging task. Most of the existing methodologies focus on English language tweets and only limited work has been reported for Hindi and German language posts. Also, the importance of feature selection methods is not explored much for this problem. In this research work, an enhanced ensemble classifier approach is proposed to identify hate and offensive content posted in Hindi or German languages. In the proposed approach, CHI square based feature selection method is combined with a Random Forest Classifier to classify the tweets. This work was submitted to Hate and Offensive Content Identification (HASOC) task@FIRE2019. From the various experiments conducted on the released HASOC dataset, it is shown that an accuracy of 81% and 64% was achieved on German and Hindi language tweets.*

Keywords: Hate Speech Identification, Ensemble Classifier, Chi Square Feature Selection, German, Hindi, Social Media

Received: 21 September 2019, Revised 18 February 2020, Accepted 5 March 2020

DOI: 10.6025/jet/2020/11/2/70-76

1. Introduction

Nowadays many people post their opinions, thoughts and comments on social websites like face book, twitter etc. due to the advanced technologies. The offensive and hate speech posted in social media increases every day and the companies are investing heavily to identify such offensive tweets. As these kind of offensive tweets contain different hash tags, emojis and follow various language styles, it is highly challenging to monitor and control such hate crimes manually.

To overcome the above issues, machine learning based methods were proposed in existing works [1, 2] and they focused on detecting common hate speech, not particular to offensive speech. Even though hate speech detection problem on English language has been studied by various researchers, only few works were reported for German and Hindi language tweets. The lexicon based and rule based approaches were followed in the existing works which is not able to generalize well. Also, traditional *tf-idf* based methods were used with simple linear classifiers and emphasis was given to other feature weighting methods. In this research work, an attempt is made to study the importance of feature selection methods along with the power of ensemble based classifiers. We have proposed an enhanced ensemble classifier with the CHI square based feature selection method to select the important features.

This research work was submitted to Hate and Offensive Content Identification (HASOC) task@FIRE2019. As part of the task, the organizers released the datasets containing the tweets in German and Hindi languages. The task is to identify the tweets that contain the hate and offensive content. To perform this binary classification task, we applied various machine learning techniques by extracting suitable features from the given data. To study the importance of feature selection methods, we conducted experiments with different feature selection methods such as TF-IDF Mutual Information and CHI square based approach. Among the two datasets, German dataset was highly imbalanced, so we have applied the widely used SMOTE analysis. To design a suitable predictive model, we conducted experiments with various machine learning techniques such as Logistic Regression, Support Vector Machine and Random Forest Classifier. From the experimental results, it is observed that the ensemble based approach is better than the individual classifiers. We have achieved an accuracy of 81% on German dataset and 65% on Hindi dataset, applying Random Forest classifier with CHI square based feature selection.

The paper is organized as follows: Related works are presented in Section 2 and the proposed methodology is detailed in Section 3. The experimental results and discussion are briefed in Section 4 followed by conclusion in Section 5.

2. Related Works

There had been many studies reported on classifying the offensive content on the web. Greevy and Smeaton [5] used SVM and bag of words to detect offensive content on web pages. They have used PRINCP corpus of 3 million words with 2 class labels namely offensive and not offensive. BOW, n-gram word sequences and POS tagged documents were used by them to represent the dataset. But they used only SVM classifier for detection and other methods were not explored. A similar approach was suggested by Warner and Hirschberg [4] using unigrams with SVM to detect offensive content of web. Hate base is and online repository of hate speech words. T. Davidson, D. Warmsley [6] had build a classifier for Hate base. They have created unigram, bigram , trigram features weighted with its TF-IDF and calculated its Part of Speech (POS) tag. They suggested linear classifiers for classifying the offensive language. But the model was biased towards the offensive language and failed to differentiate between the common place offensive language with serious hate speech. Google had developed a tool for identifying toxicity of comments between the range of 0 to 100. C. Nobata, J. Tetreault [7] had proposed annotation of hate speech versus clean speech. They have collected news and finance dataset for the binary classification of abusive and clean tweets. They have employed Vowpal Wabbit's regression model for the features obtained through n-grams, linguistic, syntactic and distributional semantics. They have compared the accuracies of all the features but they worked only on English language and did not attempt in other languages. D. Gitari [8] had further classified the tweets into strong or weak using lexicon based approaches. They have used semantic and subjectivity approach to create lexicon and use these features for a classifier. But they used rule-based classifier instead of machine learning model which lead to low precision and recall scores. Nitesh et al. [11] over-sampled the minority class through SMOTE (Synthetic Minority Over-sampling Technique), which generated new synthetic examples along the line between the minority examples and their selected nearest neighbors.

To handle multilingual queries, code mixing and code borrowing need to be differentiated. The borrowing likeliness of English words in Hindi language was determined by a novel relevant factor [14]. In this work, both Hindi and English tweets were considered to find the relevant words. Various feature weighting methods have been proposed for URL classification and sentiment analysis problems and the effectiveness of different classifiers were studied. The importance of features like tf-idf and mutual information in determining the category of a web page was explored by using URL based features [15]. For sentiment analysis on movie reviews, the tf-idf and word2vec methods were applied and the effectiveness of deep learning model has been studied in [16]. A novel feature weighting method has been proposed for Naive Bayes classifier [17], for the problem of categorizing the URLs by considering only the features derived from URLs. In this work, a variant of CHI square method was suggested to find the goodness of features and it was embedded into the calculation of likelihood probability for the Na¨1ve Bayes Classifier. Using linear SVM weights as features, URL classification was performed in [18]. These URL features were automatically learnt and data-set independent dictionary was constructed to classify the URLs. In another work [19], transfer learning approach was preferred to learn the features from Convolutional Neural Network and it has been used as input to SVM for classifying the URLs that are generated using Domain Generated Algorithms. In all the above mentioned works, the significance of feature weighting methods have been studied for classifying the web pages. GermEval is a shared task focused on offensive language identification in German tweets (8500 tweets). Wieg and et al. (2018) [21] further applied the idea to Waseem et al to this task. They experimented with detecting offensive vs. non-offensive tweets, and also with a second task on further sub-classifying the offensive tweets as, insult, abuse or profanity. The 2018Workshop on Trolling, Aggression, and Cyber bullying (TRAC) hosted a shared task focused on detecting aggressive text in both English

and Hindi [22].

The dataset from this task is available to the public and contains 15,869 Facebook comments labeled as overtly aggressive(OAG), covertly aggressive(CAG), or non-aggressive(CAG). The best-performing scores was obtained using convolutional neural networks (CNN), recurrent neural networks, and LSTM for their approach. Offensive Language Identification Dataset (OLID) dataset, which was built specifically for this task was annotated using a hierarchical three-level annotation model introduced in Zampieri et al. [20]. Three sub tasks include Offensive Language Identification (Not Offensive, Offensive), Categorization of Offensive Language (Targeted Insult, Untargeted), Offensive Language Target Identification (Individual, Group, Other) [23]. In all the above methods, the importance of determining the offensive content is emphasized.

3. Proposed Methodology

The task of identifying the hate and offensive content in the tweets is considered as a binary classification problem. The performance of any binary classifier depends on the suitable features and chosen machine learning algorithm. In this work, three different feature selection methods were chosen to viz., :i) TF-IDF(Term Frequency / Inverse Data Frequency ii) Mutual Information and iii) CHI square. Also, the effectiveness of ensemble method has been studied by applying on three classifiers viz., Logistic Regression, Support Vector Machine and Random Forest Classifier. To identify hate and offensive speech on two data sets viz., German dataset and Hindi dataset, the following steps are performed:

– Translation of tweets to English

– Pre-processing and Tokenization

– Feature Extraction by applying three variants viz., TF-IDF, Mutual Information and CHI square

– Performing SMOTE analysis (this step is required only for German dataset, as it is highly imbalanced)

– Building the model and predicting whether the given tweet is offensive or not by using the model.

3.1. Translation of Tweets

In this task, we have been provided with two different language datasets (German and Hindi). As a first step, the tweets are translated to English language. For example, a tweet in German ”Frank Rennicke – Ich binxa0stolz” was converted by employing MLtranslate and it results in the corresponding English tweet Frank Rennicke - I am proud. For this translation process, ML Translator API was used, which is a Google’s Neural Machine Translation (NMT) system [24]. This translation method was widely used because of its simplicity and zero-shot translation. Melvin et al.[24] proposed a single Neural Translation multilingual model that shares the same encoder, decoder and attention modules for all the languages without increasing the complexity of model. Also, as the parameters are shared across all the languages, it generalizes well to multiple languages. This NMT model has the advantage of zero-shot translation, as several language pairs are used in a single model and unseen word pairs in different languages were also learnt by the model. We found this translation process as suitable for this task and hence applied the same for converting the tweets in German / Hindi to English.

3.2. Pre-processing and Tokenization

Hash tags provide insights about a specific ideology by a group of people. These tags provide vital information for text classification, especially in the case of identification of offensive language in tweets. So we have processed the hash tags and obtained tokenized words out of it after segmenting the tokens. For example, after applying the hash tag segmentation on the pre-processed tweet #everyhingisgood, we obtain everything is good. Lemmatization is the process of reducing the word to its root form, which is helpful. We have used NLTK (Natural Language Tool Kit) WordNet Lemmatizer for performing lemmatization. Consider the following example, Koeln Mohamed recognizes no German right but only the #Scharia. That he wanted to break Cologne Cathedral was just a joke but when he comes out of jail, he has no more pity. After lemmatising, it becomes koeln mohamed recognizes german right scharia wanted break cologne cathedral joke come jail pity.

3.3. Feature Extraction

In any text classification task, the feature extraction plays an important role. To extract the suitable features from the pre-processed data, we have used three variants namely TF-IDF, Mutual Information and Chi-square.

TF-IDF: The TF-IDF (Term Frequency – Inverse Document Frequency) is the well-known weighting scheme and this score is

calculated based on the count of terms that are present in every tweet with the terms present in the entire corpus. As it extracts most descriptive terms from the tweet collection and simple to implement, we have chosen this feature weighting scheme, In our experiments, the minimum frequency of the word is set to 5 and maximum number of words is set to 5000.

Mutual Information: Mutual Information (MI) is the measure of dependence between two random variables, and it can be used to find the dependency between the input features and the output categories in the context of feature selection for text classification problems. For the given task of classifying the tweets, we can calculate the amount of information a particular word contributes to the class label (offensive). If the mutual information is high, then the feature has high relevance to that target and if it is zero, there is no relevance.

In the HASOC German (also for Hindi) dataset, we have calculated the values of *a, c, b* and *d* based on the number of training tweets in positive / negative category that contains / does not contain the term *ti*. The mutual information is obtained by using the formula shown below.

$$MI = log_2(max(aN/(a + c)N, cN/(a + c)N) \quad (1)$$

where '*a*' denotes the number of positive category tweets in training data that contains the term *ti* '*b*' denotes the number of positive category tweets in training data that do not contain the term *ti* '*c*' denotes the number of negative category tweets in training data that contains the term *ti* '*d*' denotes the number of negative category tweets in training data that do not contain the term *ti*.

Chi Square: The Chi-Square test is generally applied to find the relationship between two variables. The effectiveness of Chi-square based feature selection method has been reported in various text / web page classification problems [15,17]. In Natural Language Processing, identifying the relevant words is important to increase the efficiency of the classification algorithm. The Chi square statistic would be small if the term is uncorrelated with the class and would be high, if the term is correlated. In this task, we have calculated Chi-square statistic using the dataset and selected the terms with high score as they are the most informative features. Its formula is given below using the same notations *a, b, c* and *d* mentioned above.

$$Chi = (N(ad - bc)^2)/((a + c)(b + d)(a + b)(c + d)) \quad (2)$$

3.4. Addressing Imbalanced data and Classification

The German dataset was a highly imbalance dataset, that contains 3412 hate and offensive tweets with 407 non-offensive tweets, so SMOTE analysis is performed. For the Hindi dataset, this step was ignored, as it is a balanced dataset.

Random Oversampling and under sampling The mechanics of random oversampling follow naturally from its description by adding a group of *N* number of samples from the minority category. While oversampling adds data to the original data set, random under sampling removes the data from the data set. Both the methods try to alter the size of the original data set. Even though, training accuracy may increase by applying this method, the model performance will be relative low on testing data [13].

SMOTE Analysis We have applied SMOTE (Synthetic Minority Oversampling Technique) from sklearn. By this oversampling technique, the size of minority class tweets are increased to the size of majority class tweets. This method generates synthetic minority examples to over-sample the minority class. For every minority example, its *k* (which is set to 5 in SMOTE) nearest neighbours of the same class are calculated, then some examples are randomly selected from them according to the over-sampling rate. SMOTE analysis was applied to give better performance compared with other sampling Techniques [11].

Classification After making the dataset suitable for training, two different models were designed, one with Logistic Regression and another one with the ensemble classifier Random Forest by varying the feature weighting methods viz., TF-IDF, Mutual Information and Chi-square.

4. Experimental Results

To study the performance of the proposed method on the German (and Hindi) datasets, various experiments were conducted. For implementation, we used Python 3 and scikit-learn library. All the experiments were carried on a workstation with Intel-

Xeon Quad Core Processor, 32 GB RAM, NVIDIA Quadro P4000 GPU 8GB. For the initial experiments, we have divided the released training data into training set and validation set and conducted the experiments using accuracy as the performance metric. Finally the performance of the proposed system was tested on the test set released by the organizers. For these experiments, we combined all the training and validation data into a single training set and applied the algorithm. We have reported the validation accuracy and test accuracy obtained on both German dataset and Hindi dataset.

After translation and pre-processing of tweets, tokenization was performed. Then to extract the suitable features, we have applied three variants viz., TFIDF, Mutual Information and Ch-square. First, TF-IDF vectorizer (using sklearn) was used to get maximum of 10,000 features with the minimum occurrence frequency of 2 for German dataset and 5000 features for Hindi dataset. We then tried with count vectorizer (using sklearn) and calculated Mutual Information and Chi-square values for every word token using the above mentioned formulas. By this way, a total of 12,717 features were extracted for German dataset and 15,111 features were extracted for Hindi dataset. We have used the above features and used Logistic Regression (LR) and Random Forest (RF) classifier with three variants viz., TF-IDF, Mutual Information and Chi-square values. The accuracy of simple and ensemble classifiers on validation set and test set was presented in Table 1 and Table 2.

It is observed from Table 1 that, on German dataset, among the three feature weighting schemes, CHI square based feature weighting method performs better than the other two methods viz.,TF-IDF and MI with Random Forest Classifier. A validation accuracy of 90% was achieved while combining CHI square based features feature with the ensemble classifier Random Forest. It is also to be noted that, MI performs better than TF-IDF and resulted in 88% and 89% validation accuracy on German dataset with the single Logistic classifier and with Random Forest classifier. On Hindi dataset, the validation accuracy of 79% was achieved with the Random Forest classifier for the CHI square based feature selection. Based on the inference on the validation set, we have applied CHI square with

Dataset / Methods	**Logistic Regression**			**Random Forest**		
	TF-IDF	**MI**	**CHI**	**TF-IDF**	**MI**	**CHI**
German	86	88	83	88	89	90
Hindi	77	69	78	75	72	79

Table 1. Performance - Single and Ensemble classifier - Validation Accuracy

Random Forest classifier on the released test set and the results are reported in Table 2. We have obtained an accuracy of 81% and 64% on German dataset and Hindi dataset respectively.

Dataset	**Random Forest with CHI**
German	81
Hindi	64

Table 2. Performance of proposed approach - Test Accuracy

5. Conclusion

This work was submitted to the FIRE2019 task, Identification of Hate and Offensive Speech in Indo-European Languages. In this research work, the problem of identifying the hate and offensive content in tweets have been experimentally studied on two different language datasets German and Hindi. The importance of feature weighting methods was analysed by using three different variants viz., TF-IDF, Mutual Information and CHI square based feature selection. After choosing the suitable feature selection method, we have studied the significance of ensemble classifier over individual classifier. Among the released

datasets, German dataset was highly imbalanced, so we applied SMOTE analysis and then performed classification. From the experimental results, it is shown that the performance of the Random Forest classifier with CHI square based feature selection method is better than the other methods and a test accuracy of 81% and 64% were achieved on German and Hindi dataset respectively. In this work, we have restricted to machine learning approaches with suitable feature selection method and deep learning techniques will be explored in future.

6. Acknowledgment

The authors would like to thank the management of Vellore Institute of Technology, Chennai for providing the support to carry out this work. The first would like to thank the Department of Science and Engineering Research Board (SERB), Government of India for their financial grant (Award Number: ECR/2016/00484) for this research work.

References

[1] Burnap, P., Williams, M. L. (2015). Cyber hate speech on twitter: An appication of machine classification and statistical modeling for policy and decision making. *In: Policy and Internet*, Vol.7.2, p 223–242.

[2] Kwok, I., Wang, Y. (2013). Locate the hate: Detecting tweets against blacks. *In: Twenty-Seventh AAAI Conference on Artificial Intelligenc*e, p 1621-1622.

[3] de Gibert, O., Perez, N., Garc'ia-Pablos, A., Cuadros, M. (2018). Hate Speech Dataset from a White Supremacy Forum. *In: 2nd Workshop on Abusive Language Online*, p 11-20 (2018).

[4] Warner, W., Hirschberg, J. (2012). Detecting Hate Speech on the World Wide Web. *In: Proceedings of the Second Workshop on Language in Social Media*, p 19-26.

[5] Greevy, E., Smeaton, A. F. (2004). Classifying racist texts using a support vector machine. *In: Proceedings of the* 27th *annual international conference on Research and development in information retrieval SIGIR* '04, p 468 – 469.

[6] Davidson, T., Warmsley, D., Macy, M., Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *In: Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (ICWSM 2017), p 512-515.

[7] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y.,Chang, Y. (2016). Abusive Language Detection in Online User Content. In: Proceedings of the 25th International Conference on World Wide Web (WWW 2016), p 145-153.

[8] Gitari, D., Zuping, Z., Damien, H., Long, J. (2015). A Lexicon-based Ap oach for Hate Speech Detection. *In: International Journal of Multimedia and Ubiquitous Engineering*, vol.10.4, p 215-230.

[9] Hall, M., Smith L. (1998). Practical feature subset selection for machine learning. *In: Proceedings of the* 21st *Australasian Conference on Computer Science*, p 181-191.

[10] Wu, H., Gu, X. (2017). Balancing Between Over-Weighting and Under-Weighting in Supervised Term Weighting. In: *International Journal of Information Processing and Management*, vol.53, p 547-557.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. In: *Journal of Artificial Intelligence Research*, vol.16, p 321-357.

[12] Han, H., Wang, W. Y., Mao, B. H. (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *In: International Conference on Intelligent Computing*, p 878-887.

[13] He, H., Garica, E. A. (2009). Learning from imbalanced data. *In: IEEE Transactions On Knowledge and Data Engineering*, vol. 21, p 1263-1284.

[14] Rajalakshmi, R., Agrawal, R. (2017). Borrowing Likeliness Ranking based on Relevance Factor, *In: Proceedings of the Fourth ACM IKDD Conferences on Data Sciences*, CODS 2017, India, p 12:1–12:2

[15] Rajalakshmi, R., Xaviar, S. (2017). Experimental Study of Feature Weighting Techniques for URL Based Webpage Classification, *Procedia Computer Science*, Vol. 115, p 218-225.

[16] Sivakumar, S., Rajalakshmi, R. (2019). Comparative evaluation of various feature weighting methods on movie reviews, *Advances in Intelligent Systems and Computing*, Vol-711, p 721-730.

[17] Rajalakshmi, R., Aravindan, C. (2018). Naive Bayes ap oach for URL classification with supervised feature selection and rejection framework, *Computational Intelligence*, 34(1), p 363-396.

[18] Rajalakshmi, R., Aravindan, C. (2018). An Effective and Discriminative Feature Learning for URL Based Web Page Classification, *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Miyazaki, Japan, 2018, p 1374-1379.

[19] Rajalakshmi, R., Ramraj, S., Ramesh Kannan, R. (2019). Transfer learning ap oach for identification of malicious domain names, *Communications in Computer and Information Science*, Vol. 969, p 656-666.

[20] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). *In: Proceedings of the 13th International Workshop on Semantic Evaluation*, p 75-86.

[21] Wiegand, M., Siegel, M., Rup nhofer, J. (2018). Overview of the germeval 2018 shared task on the identification of offensive language.

[22] Kumar, R., Ojha, A. K., Malmasi, S., Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. *In: Proceedings of TRAC*.

[23] Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019). Predicting the Type and Target of Offensive Posts in Social Media. *In*: *Proceedings of NAACL*.

[24] Johnson, Melvin., Schuster, Mike., Le, Quoc V., Krikun, Maxim., Wu, Yonghui., Chen, Zhifeng., Thorat, Nikhil., Vi'egas, Fernanda., Wattenberg, Martin., Corrado, Greg., Hughes, Macduff., Dean, Jeffrey. (2017). Google's Multilingual Neural Machine Translation System: *Enabling Zero-Shot Translation*, Vol 5, p 339—351.

[25] Modha, S., Mandl, T., Majumder, P., Patel, D. (2019). Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo European Languages. *In*: *Proceedings of the* 11*th annual meeting of the Forum for Information Retrieval Evaluation* (December).