Inappropriate Language Detection using Machine Learning

Garvit Kumar Arya[1] Mayank Sharma[2] Aditya Jyala[3]
[1,2,3]Department of Information Technology
[1,2,3]IMSEC, Ghaziabad, U.P., India

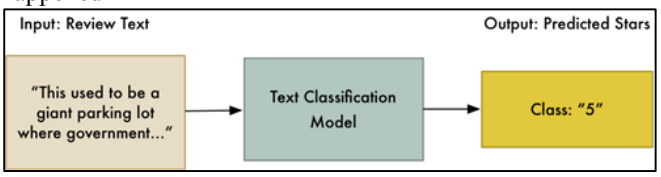
Abstract— The objective of this paper was to develop such a mechanism which can be embedded with social media platforms to stop the inappropriate content to push over the internet. when someone does their comments on a post then there are chances that the comment contains abusive, obscene, insult, threat and similar kind of content for someone which cross the boundaries of social media platforms so to tackle such a problem "Online Abusive Comment Detector" plays a massive role. It detects the content and its kind i.e abuse, insult, e.t.c and on the basis of content send a notification as a warning to the user that the comment you've done contains an inappropriate language and suggest him/her to change the language and if he/she doesn't do so then his/her account may be blocked for a limited time period.

Key words: Machine Learning

I. INTRODUCTION

Text classification is an essential component in many applications, such as sentiment analysis, news categorization, and in our research domain of interest, abusive text detection. One of the fundamental tasks in text classification is feature representation - finding appropriate approaches to represent text content. The traditional approach is based on the occurrence-model, frequency of words (e.g. BoW) in text content. This largely ignores word orders and thus the problem of capturing semantics between words stillremains. Adding extra features that are identified by experts based on specific task requirements can alleviate the drawback of traditional features. However, this takes time and human effort and introduces domain-specific dependencies into the model. One solution for feature extraction without hand-crafting is to use deep learning methods. In particular, this trend is sparked by the emergence of word embedding techniques, such as word2vec, tf-idf, and glove. Word embedding is a distributed representation at word level which has been proven to be capable of learning word semantics. To generate a distributed feature representation at the sentence level, one of the straightforward approaches is averaging the pre-trained word embeddings. However, this reduces context information such as the sensitivity of word orders, which limits semantic knowledge.

In this paper we, ve user Naive Bayes algorithm which is widely used for text classification. It works on principle "what happens next if something has already happened"



II. STATE-OF-THE-ART CLASSIFICATION

In this section, we investigate the classification and the ways in which we can solve the problem using the classification technique. In our case, the appropriate algorithm which can be used and gives the best performance on classifying text is Naive Bayes.

Selection of best algorithm needs some pre-knowledge of working of algorithms and the specific field in which they give their best performance.

A. Applications in Abuse detection

Early research in text classification on addressing abusive social media comments focused on exploring useful information such as lexical features, the user's profile and historical activities. Djuric et al. are the forerunners of implementing a neural network architecture to generate a distributed feature representation for hate speech detection. They used paragraph2vec for the modeling of comments. Compared to the BoW representation, the classification accuracy increased from 0.78 to 0.80 compared with a logistic regression classifier. Early research in text classification on addressing abusive social media comments focused on exploring useful information such as lexical features [2], the users' profile [3] and historical activities [4]. Djuric [6] et al. are the forerunners of implementing a neural network architecture to generate a distributed feature representation for hate speech detection. They used paragraph2vec [9] for the modeling of comments. Compared to the BoW representation, the classification accuracy increased from 0.78 to 0.80 compared with a logistic regression classifier. Subsequently, Nobata et al. [11] also conducted a set of comprehensive experiments to evaluate the performance of a variety of representations for abusive comments. They compared the paragraph2vec [9] to a number of feature representations including n-grams, linguistic and syntactic. Using an SVM classifier, the results indicated that using paragraph2vec to generate comment embeddings outperformed the linguistic and syntactic handcrafted features. In addition, they also show the performance of simply using an averaging strategy over the pre-trained word embeddings is better to the ngrams feature representation in most of the datasets.

Furthermore, there is an increasing number of researchers who started to work on complex deep neural networks for tackling the problem of abusive text detection. Badjatiya et al.[1] investigated CNN for hate speech detection in tweets, which significantly outperformed the traditional methods such as Logistic Regression and SVM; Gamback et al. [7] also conduct CNN architecture to classify tweets into four categories include racism, sexism, both (racism and sexism) and neither, they modified traditional CNN input with word embedding by concatenating character ngrams; Park et al.[12] proposed an improved CNN model that

combined word embeddings and character embeddings as well.

Mehdad et al. [10] implemented RNN using characters as input instead of words, which achieved an increase of approximately 8% in average class accuracy. An advanced RNN model, bi-directional LSTM with attention mechanism which adds weights for the importance of each input, was proposed by Gao et al.[8] and Del Vigna et al.[5]. Both of them achieved better performance compared to the one-directional LSTM. In addition, Pavlopoulos et al.[13] also showed the attention mechanism improves the performance of the RNN model when dealing with abusive comments in the Greek language.

III. METHODOLOGY

Natural Language Processing is used to remove redundancy and stopwords from the dataset because if stopwords are present in the dataset then there are chances of giving undesired result by the algorithm. There are wide applications of NLP from which we've used the TF-IDF algorithm to process and convert words into vectors because machine learning algorithms work on vectors i.e numbers instead of text i.e word.

TF-IDF is a well-known algorithm used to convert words into vectors to perform further operations on data. TF-IDF works on term frequency and inverse document frequency in which term frequency is the count of a specific word in a document and inverse document frequency is the log of corpus up to total documents contains the specific word.

$$\text{TF-IDF}_{w\,id} = \text{TF}_w * \log(N / \text{DF}_W)$$

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other. Inside Naive Bayes, Multinomial Naive Bayes playing the actual game, here is the formula of Naive Bayes.

$$P(c \mid x) = P(x \mid c) * P(c) / P(x)$$

where P(c|x) is Posteroir Probability, P(x|c) is the likelihood, P(c) is class prior probability and P(x) is Predictor Prior Probability.

IV. CONCLUSIONS & FUTURE WORK

The purposes of this paper were two-fold: (1) to investigate how off-the-shelf classification have been used across the two tasks within text classification - feature representation and the classification tasks itself and (2) to run preliminary experiments in abusive detection across social media datasets. We highlight the supervised approaches for classification. Secondly, we attempted to compare the classification performance of using traditional BOW algorithm and then TF-IDF. Due to its shallow architecture, the performance is unexceptional. We will validate this assumption in our subsequent experiments. The ultimate goal of our research is to develop a powerful classification model that can assist social media moderators to detect abusive comments efficiently and effectively. The path to the goal can be divided into two directions, designing appropriate features for abusive comments and designing an outstanding model for detection. In future work, we will focus our research in exploring deep neural networks in unsupervised mode and supervised mode.

REFERENCES

[1] Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760 (2017)
[2] Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting offensive language in social media to protect adolescent online safety. In: Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). pp. 71–80. IEEE (2012)
[3] Dadvar, M., de Jong, F.M., Ordelman, R., Trieschnigg, R.: Improved cyberbullying detection using gender information (2012)
[4] Dadvar, M., Trieschnigg, D., Ordelman, R., de Jong, F.: Improving cyberbullying detection with user context. In: ECIR. pp. 693–696. Springer (2013)
[5] Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., Tesconi, M.: Hate me, hate me not: Hate speech detection on facebook (2017)
[6] Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. pp. 29–30. ACM (2015)
[7] Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online. pp. 85–90 (2017)
[8] Gao, L., Huang, R.: Detecting online hate speech using context aware models (2017)14. Hill, F., Cho, K., Korhonen, A.: Learning distributed representations of sentences from unlabelled data. arXiv preprint arXiv:1602.03483 (2016)
[9] Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st ICML-14. pp. 1188–1196 (2014)
[10] Mehdad, Y., Tetreault, J.R.: Do characters abuse more than words? In: SIGDIAL Conference. pp. 299–303 (2016)
[11] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th International Conference on World Wide Web. pp. 145–153 (2016)
[12] Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206 (2017)
[13] Pavlopoulos, J., Malakasiotis, P., Androutsopoulos, I.: Deep learning for user com-ment moderation. arXiv preprint arXiv:1705.09993 (2017)

