Towards Developing Effective Machine Learning Frameworks to Identify Toxic Conversations Over Social Media

Ritam Majumder and Tanmay Basu
Department of Computer Science
Ramakrishna Mission Vivekakanda Educational and Research Institute
Belur Math, Howrah, West Bengal, India

Abstract—The advent of social media has great impact in the society and huge number of conversations are being made over social media everyday. Unfortunately, many of these conversations are meant for personal attacks or include abusive comments, which may have adverse effect to particular communities or individuals. The same may raise doubts about the liability and popularity of the social media forums, which should be prevented. A set of machine learning classifiers have been explored here to automatically identify abusive or toxic comments from social media posts. The empirical analysis on a data set of Wikipedia Talk Page using different such classifiers including a recurrent neural network has shown significant improvement towards this direction.

Keywords: toxic comment classification, deep learning, text categorization

1. Introduction

The impact of the comments posted by the users in online discussion forums have significant effects toward their popularity. These comments may be harsh or abusive and are also known as toxic comments. Toxic comments are defined as a rude, disrespectful, or unreasonable comments that are likely to make the other users to leave a discussion [1]. The adverse effects of toxic comments over social media and news portals should be prevented. According to Pew report [2], it says 73% of internet users who are adult, have seen someone harassed online and 40% users have experienced it personally [3]. Therefore the online forums employ moderators to identify toxic comments and subsequently prevent them from posting. However, this task is manual, labor intensive and time consuming as the volume of the data is huge. Therefore, the process should be automated.

Many such tools using different machine learning classifiers have been developed to categorize the toxic and non toxic comments from social media posts [3] [4], which is a two class classification problem. Pavlopoulos et al. have proposed a supervised recurrent neural network (RNN) based method for moderation of user comments over social media [4]. Wulczyn et al. have introduced a method by combining crowdsourcing and machine learning to analyze personal attacks over online conversations and they have performed empirical analysis using ROC curve on a wikipedia data set. Conventionally, the toxic comments contain slang terms. In practice, comments may not contain slang e.g., consider the following comment - You have literally no potential to accomplish this significant job. This is abusive in nature, however, it does not contain any harsh word or slang. On the other hand consider another comment from Wikipedia Talk Page [5] - You guys are so stupid. It makes me wanna kill myself; where stupid is certainly an abusive word. However, this is not a toxic comment and it seems that it has been posted to have fun. The challenge here is to identify the toxicity in terms of the context of the words which is pretty difficult.

The challenges to build an effective method for identifying toxic comments of multiple categories is addressed in this article. The toxic comments considered here are divided into six categories, namely, toxic, severe toxic, obscene, insult, threat and identity hate. Moreover, a comment can belong to one class or multiple classes. Consider the following two examples from Wikipedia Talk Page [5].

- He thinks they are copyright violation. Pity him, he is sick!
- It was very constructive you are just very very stupid

The first one belongs to toxic category, whereas the other one belongs to toxic, obscene and insult categories. Note that the types of toxicity in a comment may not follow the regular way of communication. Consequently, it leads to a difficult situation to find such class overlaps in high dimensional and sparse corpora.

In this work, we have have explored different machine learning classifiers and feature engineering schemes to effectively identify different types of toxicity in the corpus. We have used a corpus of Wikipedia Talk Page for experimental analysis. The corpus has been released by Kaggle [5] as part of a challenge regarding toxic comment classification. The corpus has 150k comments in the training set, which has been used to train a machine learning classifier. The corpus is highly unbalanced i.e., out of these 150k comments only 16k comments are toxic and these 16k comments are divided into six toxic categories discussed above. Moreover, the toxic categories are highly overlapped as mentioned earlier. The test set has 150k comments, which has been used to eval-
tuate the performance of different classifiers used here. The empirical analysis using ROC area under the curve [3] has shown that a combination of RNN and convolutional neural network (CNN) with GLOVE embeddings [4] outperforms the other state of the art classifiers.

2. Experimental Evaluation

The performance of different state of the arts machine learning classifiers has been studied in this work to effectively identify toxicity of comments belong to multiple categories. The corpus has been released as part of a Kaggle challenge. The performance of a model has been evaluated using ROC area under the curve [3]. The actual class labels of the comments of the test set has not yet released by Kaggle [5]. However, the proposed model can be communicated to Kaggle submission platform [5] to check the performance of the model.

Initially, we have removed all the English stopwords following a standard list in NLTK1, a natural language toolkit. Therefore, the bag of word n-grams (\( n \leq 3 \)), i.e., each comment becomes a bag containing 1 term, 2 terms and 3 terms) [4] are considered to create the term-document matrix. Subsequently, we have applied Random Forest (RF) [4], Support Vector Machine (SVM) [4] and Logistic Regression (LR) [4] classifiers and tuned different parameters (e.g., penalty parameter of svm) of these classifiers using 10-fold cross validation on the training set. We have also used a combined model of RNN and CNN on the training set, where we have applied only Gated Recurrent Unit (GRU) as the recurrent layer [6] and then applied the best model on the test set. Table 1 shows that Logistic Regression using word \( n \)-grams outperforms the other classifiers. Note that the abusive words are mostly noun or adjective or verb, if not misspelt. Therefore we have extracted these parts of speech using NLTK parts of speech tagger2. It may be noted that LR with word \( n \)-grams has shown best result in Table 1. Hence we have created a vocabulary of nouns, adjectives and verbs from the given corpus and then implemented LR using word \( n \)-grams on this vocabulary. The ROC AUC of this model is 0.9794, which is less than the ROC AUC of LR using the original vocabulary as shown in Table 1. This indicates that only noun, adjective and verbs are not enough to extract potential information regarding toxicity. We have discarded the delimiters when n-grams have shown best performance on Wiki News and it is created by Facebook [8].

we considered bag of words as features. However, certain delimeters e.g., exclamation marks (!) may be significant in identifying toxicity. Therefore we have created the term document matrix using both word \( n \)-grams and character \( n \)-grams (\( n \leq 5 \), each comment becomes a bag containing 1 character, 2 characters and so on) [4] and subsequently LR classifier is implemented. The ROC AUC of this model is 0.9804, which is a little improvement over LR using only word \( n \)-grams. We have also studied the performance of neural network models using different types of pretrained word embeddings [4]. GloVe embeddings [7] is a pre-trained word embeddings on 840B tokens, 2.2M vocabulary and 300 dimensional vectors and it is developed by the natural language processing group at Stanford. Fasttext embeddings is pre-trained on 2 million word vectors with 600 billion tokens over Wiki News and it is created by Facebook [8]. The performance of the combined model of RNN and CNN using Glove and Fasttext embeddings are also studied. Table 2 shows that the results of these models perform better than the earlier models implemented in this article. Moreover, the combined model of RNN and CNN using Glove embedding performs better than the same using Fasttext embeddings.

3. CONCLUSIONS

We have studied the performance of different state of the art classifiers to identify toxic comments of multiple overlapped categories. It has been observed from the results that the neural network models using different types word embeddings effectively accomplish the task.

Table 2: State of the art classifiers using Word n-grams

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>0.9432</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.7531</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.9800</td>
</tr>
<tr>
<td>RNN and CNN</td>
<td>0.9213</td>
</tr>
</tbody>
</table>

1http://www.nltk.org/nltk_data/
2https://www.nltk.org/api/nltk.tag.html

References