Toxic Comment Classification

Anirudh Sagar Gollapalli (EE18B007), Jayadev Joy (EE18B011)
Naga Deepthi Chimbili (EE18B043), Varshitha Vadapally (EE18B065)

Department of Electrical Engineering, IIT Madras

April 12, 2021

Abstract

The anonymity that various social media services provide for their users, unfortunately brings out the worst in some people, which for most instances manifests itself as"toxic comments". This has attracted the attention of the creators of various social media platforms and websites. Moderating this manually is exceedingly impractical, given the tremendous pace at which the usage of social media services is increasing. Hence, a lot of social media giants automated this process of detecting toxic comments by implementing Machine Learning (ML) Algorithms.There are different types of toxicity: Toxic, Severely toxic, Obscene, Threat, Insult, Identity hate. What we attempt to do is to build a model which predicts a probability of each type of toxicity for each comment. This is a multi-label classification where one comment can belong to more than one label simultaneously. For example, a comment maybe toxic, obscene and insulting at the same time. It may also happen that the comment is non-toxic and hence does not belong to any of the six labels.

The Conversation AI Team, a research initiative, founded by Jigsaw and Google put forth a data-set for this purpose by curating a large number of Wikipedia comments, which we are going to use for our classifier.

1 Background and Motivation

The internet is growing at an exponential rate as one of the largest platforms for human conversations. In order to enforce healthy conversations, restriction of negative interactions and behavior is necessary. Toxic comments fall under this category. Hence, there is a need for a solution that can be implemented on a large scale operation, which is where manual moderation fails. A solution for this problem is to implement a classification module for the various types of toxic comments. This can be used for automatic recognition of toxic comments which is very useful for moderators as well as for users who would want to filter unwanted contents. With all the progress and improvement in IT and data science, there is a requirement of a properly designed technique to find and

isolate these kinds of comments that we call toxic, which is what motivated us to undertake this project.

2 Data-set description

The Data-set used for this task is sourced from a Kaggle competition and is split into training data and test data. It is composed of comments from Wikipedia's talk page edits.

The training data-set consists of total 159571 instances with comments and corresponding multiple binomial labels:

- Severely Toxic
- Toxic
- Obscene
- Threat
- Insult
- Identity Hate

Sample instances of the data-set are shown below in the figure:

	A	B	C	D	E	F	G	H
1	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
2	0000997932d777bf	Explanation	0	0	0	0	0	0
3	000103f0d9cfb60f	Shut up	1	0	1	0	1	0
4	000113f07ec002fd	Hey man, I'm really not trying to edit war. It's just t	0	0	0	0	0	0

Figure 1: Sample data-set

Our task is to predict the probability of each type of toxicity for each comment.

3 Existing Work

Many researchers have attempted to find a good and efficient solution for this problem. Methods to solve this problem in the recent past were limited to Naive Bayes, Random Forest, Decision Tree. But, the recent advancements made in Machine Learning gave rise to much more efficient and complex methods. And so, new algorithms such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), etc. are being used now. Here is a list of papers of the approaches they used :

1. *Multilabel Toxic Comment Classification Using Supervised Machine Learning Algorithms.* Darshin Kalpesh Shah, Meet Ashok Sanghvi, Raj Paresh Mehta, Prasham Sanjay Shah, Artika Singh

'The aim of this paper is to perform multi-label text categorization, where each comment could belong to multiple toxic labels at the same time. We tested two models: RNN and LSTM. Their performance is significantly better than that of Logistic Regression and ExtraTrees, which are baseline models.'

2. *Detecting Abusive Comments Using Ensemble Deep Learning Algorithms.* Ravinder Ahuja, Alisha Banga, S C Sharma

   'We have applied four classification algorithms: Naive Bayes, Random Forest, Decision Tree, and Support Vector Machine, with Bag of Words features. Deep learning algorithms: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and an ensemble of LSTM and CNN are applied using GloVe and fastText word embedding'

3. *Toxic Comment Classification.* Sara Zaheri, Jeff Leath, David Stroud

   'Accordingly, aiming to find and develop an efficient algorithm to identify toxic comments, the current study implemented several algorithms, including Naıve Bayes (as a benchmark), Recurrent Neural Network (RNN), and Long Short Term Memory (LSTM).'

4 Proposed Work

As this project involves text based data to train and test, first we would have to preprocess the data and then use that to train the model. First, we would test models like Naive Bayes, Decision Tree. Since these are baseline models, we can't expect them to provide high accuracy. So, we would also like to test models like Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) and try to use methods like Word embedding, Hyper-parameter tuning, etc to improve the accuracy and find the best model based on it.

5 References

1. Navoneel Chakrabarty. *A Machine Learning Approach to Comment Toxicity Classification.* Jalpaiguri Government Engineering College, West Bengal, India.

2. Darko Androcec. *Machine learning methods for toxic comment classification: A systematic review.* University of Zagreb Pavlinska, Croatia.

3. Sara Zaheri, Jeff Leath, David Stroud. *SMU Data Science Review - Toxic Comment Classification.* Southern Methodist University.

4. Deepan Das. *Toxic Comment Classification.* University of Wisconsin, Madison.