Toxic comment classification: Identify and classify toxic online comments

By Lesi Wei_a1797017

Faculty of Engineering, Computer and Mathematical Sciences, The University of Adelaide

Report submitted for COMP SCI 7209 Big Data Analysis and Project at the Faculty of Engineering, Computer and Mathematical Sciences, University of Adelaide towards the Master of Data Science

Abstract

Using the neural network model with long short-term memory and 1demesional convolutional neural network layer to achieve an accuracy score of 0.97050 to the toxic comments classification challenge problems. Also building the naïve bayes and logistic regression model as baseline model to be the reference.

Introduction

The internet became a such important part of life in the modern society. We take the internet real serious nowadays, according to the Digital 2019 report, on average people spent 6 hour and 42 minutes on the internet, which is more than a quarter of time for a day. There are 5.11 billion unique mobile users in the world today, and there are 4.39 billion internet users in 2019, 3.48 billion social media users. [1] Nowadays, the flow of data over the web has grown up dramatically, particularly with the looks of social networking sites. thanks to this, a very important task now could be the development of algorithms to mechanically classify the social networks content as "positive" or "negative", to stop potential hurt to the society. In the recent years the authorities start using social networks as resource to detect toxic or threat and arrest the person who make those comments. In San Francisco, a person who name is Peralta, a 35-year-old activist and musician, didn't think twice about the 23 January Facebook thread until two months later, when he learned that police had issued a warrant for his arrest – accusing him of threatening to kill law enforcement.[2] So there are high demand for pre-processing the nature language and give them label to which it can be category to different types. In the paper, the main purpose is using machine learning algorithm to classify the toxic comments. The task is solved on the example of data of the site Kaggle.com(machine learning site: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 16.03.2018).

Text Classification Problem

As the human we are good at identify the text and image, when we look or read test, we can easily to identify the meaning of those words, but if it is large scale of number that's very hard for human to calculated the answer. However, computers are good at number and calculations. therefor any kinds of text information must convert into the language that computers can read in order to analyze the meaning behind of those words. So before implement any machine learning and deep learning algorithm, looking at the general dataset summary is needed.

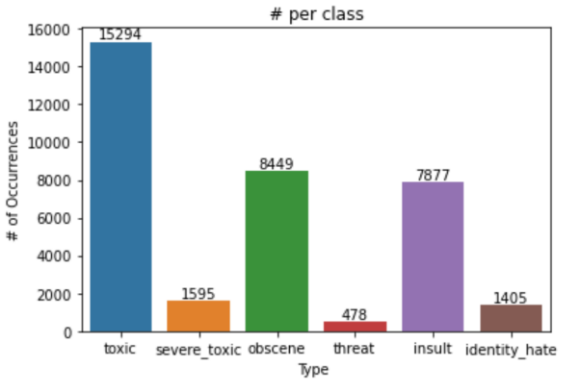

Figure 1. Summary of toxic comments

The figure above is showing that number of total toxic comments in the training dataset. There are 159571 observations in the training dataset, and there are total 143346 clean comments. So that there are total 16225 comments have at least one of

the toxic labels which about 10% of total training data. From the information above, it is easy to see that the toxic, obscene, and insult are taking majority of the toxic comments.

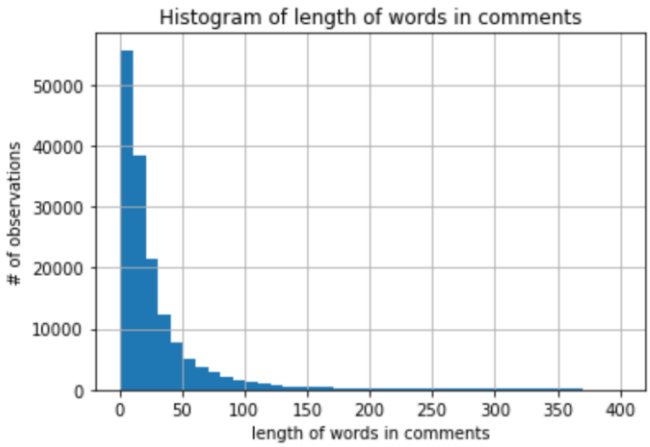

Figure 2: Histogram of words lengths

The figure above shows that the histogram of sentences length in each comment. It indicates that majority of comments are less than 200 words, and more than 95% of comments are less than 100 words in the sentences. Which this information can be used in the later model building stage.
Then the data cleaning methods being implemented to both training and testing set. First, by using the in-build function to lower all letter into lower case, then create a function that remove all website links and non-ASCI characters. After that create a function to remove all the punctuation in the comments. And some of the words in the comments are begin with "\n", so a function is being created to remove all of that. Finally, create a function that remove all the stop words in the comments which will improve the accuracy of the model since stop words are not major contribution for the meaning behind the comments.

Naïve bayes and logistic regression Models

Naive bayes classifier is a supervised learning method, which assumes that the value of an attribute influences a given class and is independent of other attribute values, but the restriction is strong and often cannot be met in real case. The naive Bayes classifier has achieved great success, showing high precision and high efficiency, with the smallest misclassification rate, and the characteristics of low time and overhead. In the training dataset, the TfidfVectorizer is being used with maximum feature of 20000 to vectorize the text into computer readable language, then building the model, and using the test set to predict the result, turns out that the average auc roc accuracy score 0.6079.

Logistic regression is one of the most used classification methods in machine learning. This is mainly because its essence is a linear model plus a mapping function

Sigmoid, which maps the continuous results obtained by the linear model to the discrete type. It is often used for binary classification problems, and in multi-classification problems it always uses the method called SoftMax. To built the logistic regression model the TfidfVectorizer still being used with maximum feature of 20000. Then building the model use the training data, and using the testing set to predict the result. The average auc roc accuracy score for the logistic regression is 0.9776.

Neural Network Model

Neural network is a machine learning model. It is an algorithmic mathematical model that imitates the behavioral characteristics of animal neural networks and performs distributed parallel information processing. This kind of network relies on the complexity of the system and achieves the purpose of processing information by adjusting the interconnection between a large number of internal nodes. The type of neural network using is the long short term memory network and one-dimensional convolutional neural network. The long short term memory network is a time recurrent neural network, suitable for processing and predicting important events with relatively long intervals and delays in time series. For the neural network model, the tokenizer methods is being used to process the text into a list of subscripts of words in a dictionary, counting from 1. And the tokenizer actually only generates a dictionary, and counts the word frequency and other information, which does not convert the text into the required vector representation. Then the model is being created with tokenized words.

TABLE 1. Comparing the validation results of the neural network

	Accuracy	Validation loss	Validation accuracy
LSTM	0.9893	0.0543	0.9877
1D-CONV	0.9877	0.0734	0.9882
LSTM+1DCONV	0.9938	0.0582	0.9934

Layer (type)	Output Shape	Param #
input_4 (InputLayer)	[(None, 200)]	0
embedding_3 (Embedding)	(None, 200, 50)	1000000
bidirectional_3 (Bidirection	(None, 200, 128)	58880
dense_5 (Dense)	(None, 200, 64)	8256
global_max_pooling1d_3 (Glob	(None, 64)	0
dropout_3 (Dropout)	(None, 64)	0
dense_6 (Dense)	(None, 6)	390

Total params: 1,067,526
Trainable params: 1,067,526
Non-trainable params: 0

Figure 3:LSTM

Layer (type)	Output Shape	Param #
input_6 (InputLayer)	[(None, 200)]	0
embedding_5 (Embedding)	(None, 200, 50)	1000000
conv1d (Conv1D)	(None, 198, 64)	9664
dense_7 (Dense)	(None, 198, 64)	4160
global_max_pooling1d_4 (Glob	(None, 64)	0
dropout_4 (Dropout)	(None, 64)	0
dense_8 (Dense)	(None, 6)	390

Total params: 1,014,214
Trainable params: 1,014,214
Non-trainable params: 0

Figure 4:1D CONV

There are three types of model is being build, with the three model with its accuracy in the table above, as it is clear to see that the model with long short term memory and 1D convolutional layer perform the best for the accuracy and the validation accuracy.

Then using the LSTM and 1d convolutional layer combine model to try out different parameter see the improving of the accuracy and decrease of the validation loss.
The graph below showing that the experiment that trying different batch and see how the different batch size effect on the validation loss. The choice of batch size 128 or 256 seems the better option for the model with better accuracy. So, the batch size 128 being chosen for the model and carry on for other parameters.

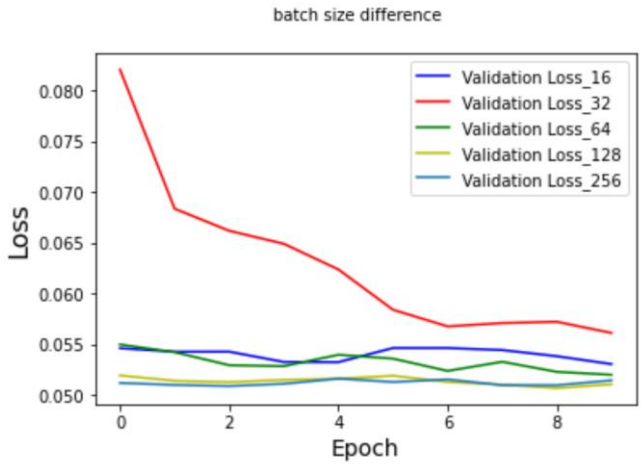

Figure 5: batch size difference compares with the validation loss

Then the following graph shows that how is the drop our rate effect the validation loss, which will cause the accuracy change. As it can see that with the lowest drop our rate 0.1 is perform better compare with other drop our rates. And the drop will update to the model for prediction and model building.

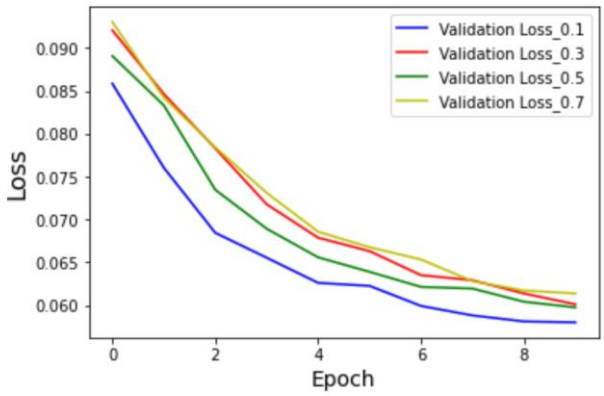

Figure 6: drop out rate difference compares with the validation loss

Then using the selected parameters above to refit the model. Then given the summary of loss and accuracy below, which at the best can get accuracy of 0.9942 and validation accuracy of 0.9940.

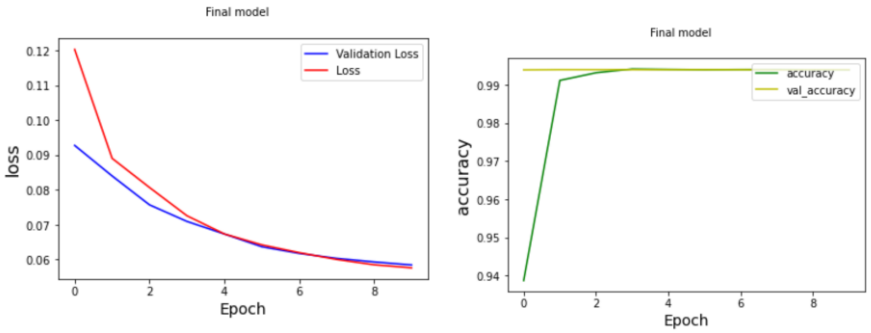

Figure 7: final model with loss and accuracy

There are few things to be consider when evaluated the model. The testing accuracy is the major determine metric when classify the toxic comments when given the model unknown comments which can determine the label of the comments. Before change the parameter it can achieve an accuracy score (column wise average roc auc) of 0.96222, then after the changing the parameter it can get to 0.97050. And the validation set is taking up the 10% of the dataset which is a good predictor for a model accuracy in the testing set. There still further improvement can be made, to try out the different methods to vectorizer transform the text information to computer readable dictionary or vector and try out different layer of neural network model.

Reference

1. Kemp, S. (2019, March 04). Digital trends 2019: Every single stat you need to know about the internet. Retrieved October 28, 2020, from https://thenextweb.com/contributors/2019/01/30/digital-trends-2019-every-single-stat-you-need-to-know-about-the-internet/

2. Levin, S. (2017, May 18). Jailed for a Facebook post: How US police target critics with arrest and prosecution. Retrieved October 28, 2020, from https://www.theguardian.com/us-news/2017/may/18/facebook-comments-arrest-prosecution

3. Toxic Comment Classification Challenge. (n.d.). Retrieved October 28, 2020, from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data