4.1 experimental settings

4.1.1 dataset statistics details

The experimental dataset used in this paper was adopted from a Kaggle competition called "toxic comment classification challenge" [1] . The entire dataset was divided into two parts, a test set and a training set, containing 159571 and 153164 comments posted on the Wikipedia talk page and the ids of the posters, respectively. The training set is divided into two parts: the validation set and the training set. A preview of the training and test sets is shown in the figure below.

表格

中度可信度描述已自动生成

Figure 8 Samples of training set

表格

中度可信度描述已自动生成

Figure 9 Samples of test set

In the training set, comments ruled as malicious or other violations were labelled as 1 in the category in which they were judged, and normal comments were labelled as 0. Comments judged to be malicious accounted for roughly 10% of the overall dataset. The content structure of the dataset is shown in the figure below.

| **Type** | **train set** | **validation set** | **Total** |
| --- | --- | --- | --- |
| normal | 129049 | 14297 | 143346 |
| toxic | 5078 | 588 | 5666 |
| severe toxic | 1446 | 149 | 1595 |
| obscene | 7587 | 862 | 8449 |
| threat | 442 | 36 | 478 |
| insult | 7065 | 812 | 7877 |
| identity hate | 1268 | 137 | 1405 |

Table 1 Dataset Structure

A typical example of each type of malicious comments in the dataset is shown in the figure below. As can be seen from the example, the categories of violation include severe toxic, obscene, threat, insult and identity hate. In the dataset, these different types of offending comments are labelled 1 for the specific category to which they belong and 0 for the rest of the labels.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type** | severe toxic | obscene | threat | insult | identity hate |
| **Samples of toxic data** | Die you piece of ordure You do not deserve to be on this planet. 137.205.183.70 | “Blocked by Freestyle frappe. Yeah, that's right, blocked for saying "IMHO WMC is a jerk." Personal insult, huh? Sock puppet? I'm fucking steamed, Freestyle frappe. And coming right after WMC's admin confirmation? If you're trying to punish me this is some of the most ridiculous bullshit gamesmanship, I've ever seen | I’ll kill you all!!!! | I'm just getting recent pics of somebody & an uneducated dufus came & removed my recent pics so help me get recent pics & try reporting that dufus & that idiot's user is Lil crazy thing. User:Pic Business | The conclusion is obvious - when you don't like what someone says because it's true, do what the Nazis did and silence that person. How Jewish of you! 12.176.152.194 |

Table 2 Samples of Malicious Comments

4.1.2 evaluation metrics

This subsection discusses the performance evaluation of the proposed model. The metrics used to evaluate the proposed model are accuracy and loss. From this subsection, the evaluation of how well the model will perform will be demonstrated.

Certain criteria are used to evaluate the performance for each model in this experiment, namely the Precision(P), the Recall (R), the harmonic mean of Precision and Recall (F1-score). After calculating these criteria for each category, macro-F1, Accuracy and Area under Curve (AUC) are used to evaluate the performance of the whole model.

In particular, P represents the number of samples with positive predictions that are predicted correctly, and R refers to how many of the positive samples in the dataset are predicted correctly. The confusion matrix representing the actual and predicted values is shown in the table below.

|  |  |  |
| --- | --- | --- |
| Actual  Predict | Positive | Negative |
| Positive | True Positive (TP) | False Negative (FN) |
| Negative | False Positive (FP) | True Negative (TN) |

Table 4 Confusion Matrix

In the table, TP: samples that are positive and predicted to be positive; FN: samples that are positive but predicted to be negative; FP: samples that are negative but predicted to be positive; TN: samples that are negative and predicted to be negative.

Therefore, the expressions of the evaluation criterion of the model are as (1) to (5):

4.1.3 baseline models

The following baseline methods for text classification are benchmarked in this study. They are useful methods that have produced positive outcomes in text classification:

GloVe + GRU: A Gated Recurrent Unit model using GloVe to perform word vectorization[2].

Transformers: a model structure that completely hinges on an attention mechanism, forgoing recurrence, to identify global dependencies between inputs and outcomes[3].

LSTM + NB-SVM: A Long Short-Term Memory model which combines Support Vector Machine builds on naive bayes features[4].

Naïve Bayes + Logistic Regression: Logistic Regression model building on naive bayes features.

Random Forest: An algorithm integrates multiple decision trees into a forest, where each decision tree is a classifier, and the random forest will aggregate all the classification votes to predict the outcome[5].

LSTM: Long Short-Term Memory.

GRU: Gated Recurrent Unit.

| **Model** | **Accuracy** | **Reported In** |
| --- | --- | --- |
| GloVe + GRU | 97.97% | Pratap D[6] |
| Transformers | 98.34% | Afzal B[7] |
| LSTM + NB-SVM | 98.10% | Howard J[8] |
| Naive Bayes + Logistic Regression | 97.95% | Abhi[9] |
| Random Forest | 96.62% | Zhu D[8] |
| LSTM | 99.16% | Project Result |
| GRU | 97.73% | Project Result |

Table 5 Baseline Model

4.1.4 parameter settings

The parameters used in this paper are shown in the table below. The dimension of the word vector obtained after the embedding layer is 13, the epochs are set to 2, the hidden unit size is 50 for both LSTM and GRU, and the batch size is 32. To prevent overfitting of the model, a dropout layer is added after the LSTM and GRU layers respectively, and the dropout rate is 0.5.

| **Parameter Setting** | **LSTM** | **GRU** |
| --- | --- | --- |
| Word Vector | 13 | 13 |
| Unit | 50 | 50 |
| Epoch | 2 | 2 |
| Batch Size | 32 | 32 |
| Dropout Rate | 0.5 | 0.5 |

Table 6 Parameter Settings

4.2 performance comparison

The evaluation values obtained on the Wikipedia toxic comments dataset for the two models in this paper, LSTM and GRU, are shown below. Since the type of classification for comments in this project is a combination of several binary classification problems, the evaluation of the model performance is separated into evaluations of the ability to distinguishing different categories.

LSTM model result evaluations for predicting different categories are shown. Overall, the LSTM model can achieve an accuracy of 91.97% in predicting whether a text is an offending comment, and an average accuracy of 97.72% in determining the category of text.



Figure 10 Evaluation of toxicity of LSTM

图形用户界面, 应用程序, Teams, 树状图

描述已自动生成

Figure 11 Confusion Metrix of toxicity of LSTM

图表, 树状图

描述已自动生成

Figure 12 F1 Metrix of toxicity



Figure 13 Evaluation of severe toxic of LSTM

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 14 Confusion Metrix of severe toxic

图表, 树状图

描述已自动生成

Figure 15 F1 Matrix of severe toxic



Figure 16 Evaluation of obscene of LSTM

图形用户界面, 应用程序, Teams, 树状图

描述已自动生成

Figure 17 Confusion Metrix of obscene

图表, 树状图

描述已自动生成

Figure 18 F1 Matrix of obscene



Figure 19 Evaluation of threat of LSTM

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 20 Confusion Metrix of threat

图表, 瀑布图

描述已自动生成

Figure 21 F1 Matrix of threat



Figure 22 Evaluation of insult of LSTM

图形用户界面, 应用程序, Teams, 树状图

描述已自动生成

Figure 23 Confusion Metrix of insult

图表, 树状图

描述已自动生成

Figure 24 F1 Matrix of insult

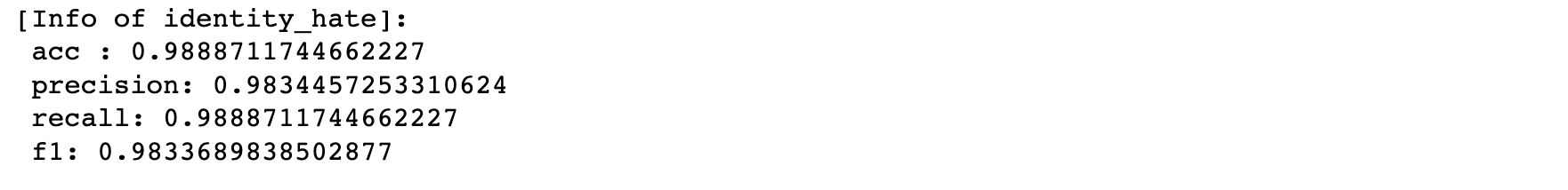


Figure 25 Evaluation of identity hate of LSTM

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 26 Confusion Metrix of identity hate

图表

描述已自动生成

Figure 27 F1 Matrix of identity hate

The following is the evaluation value of the GRU model for predicting text types. GRU model can predict whether a text is an offending comment on the accuracy of 92.78%, and the accuracy of determining the text category reached 97.83% on average.



Figure 28 Evaluation value of toxicity of GRU

图形用户界面, 应用程序, Teams, 树状图

描述已自动生成

Figure 29 Confusion Metrix of toxicity

图表, 树状图

描述已自动生成

Figure 30 F1 Matrix of toxicity



Figure 31 Evaluation value of severe toxic of GRU

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 32 Confusion Metrix of severe toxic

图表, 树状图

描述已自动生成

Figure 33 F1 Matrix of severe toxic



Figure 34 Evaluation value of threat of GRU

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 35 Confusion Metrix of threat

图表, 瀑布图, 树状图

描述已自动生成

Figure 36 F1 Matrix of threat



Figure 37 Evaluation value of insult of GRU

图形用户界面, 应用程序, Teams, 树状图

描述已自动生成

Figure 38 Confusion Metrix of insult

图表, 树状图

描述已自动生成

Figure 39 F1 Matrix of insult



Figure 40 Evaluation value of identity hate of GRU

图形用户界面, 应用程序, Teams

描述已自动生成

Figure 41 Confusion Metrix of identity hate

图表, 瀑布图

描述已自动生成

Figure 42 F1 Matrix of identity hate

The following table summarizes the performance of the LSTM model and the GRU model with the same hyperparameters setting.

| **Model** | **Category** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- | --- |
| LSTM | Toxicity | 91.97% | 90.67% | 91.97% | 90.28% |
| Severe Toxic | 99.35% | 99.04% | 99.34% | 99.16% |
| Obscene | 95.34% | 94.61% | 95.34% | 94.33% |
| Threat | 99.67% | 99.34% | 99.67% | 99.50% |
| Insult | 95.36% | 94.42% | 95.36% | 94.30% |
| Identity Hate | 98.88% | 98.34% | 98.88% | 98.33% |
| GRU | Toxicity | 92.78% | 92.44% | 92.78% | 92.59% |
| Severe Toxic | 99.32% | 99.16% | 99.32% | 99.23% |
| Obscene | 95.71% | 95.71% | 95.71% | 95.71% |
| Threat | 99.67% | 99.34% | 99.67% | 99.50% |
| Insult | 95.56% | 95.16% | 95.56% | 95.31% |
| Identity Hate | 98.88% | 97.78% | 98.88% | 98.33% |

Table 7 Model Comparison

From the table, the performance of GRU model in determining whether the text is toxic and in determining the violated text category is significantly improved compared with the LSTM model. In summary, the Accuracy of GRU can be improved by 0.81%, Precision by 0.53%, Recall by 0.81%, and F1-Score by 0.80% with the same hyperparameters.