COMP90049 Toxicity comments classification Report

Anonymous

1. Introduction

With the advancement of technology and society, online social events become more and more frequent. People are enjoying the convenience of the Internet to communicate others all over the world. However, some individuals misuse the concealment of the network to deliberately attack others [2]. As one of the most important powerful tool, automatic filtering toxic comments using machine learning is also becomeing more and more common in Internet companies. Unfortunately, due to various number of discriminatory acts in reality and biased training dataset, comments of people of different identites may be more likely to be detected as toxic. In this work, I will study the perfomance of common classification algorithm to detect toxicity of comments for different identites. The questions that should be answered are "How well do models performance?" and "If models work equally well for different identites?".

The data set used to train and predict is derived form the resource [1], specifically the one that is computed with a pre-trained language model, called the Sentence Transformer [3]. The initial data set is manually annotated by up to 10 people. The frequency of same label or identity is counted to generate a fracion score showing the probability. To simplify the task, the value of toxicity and identites are transformed to either 0 or 1 using 0.5 threshold. The whole labelled data set is separated to three files, which are "train_embedding.csv", "dev_embedding.csv" and "test_mebedding.csv". As the name indicates, the "train_embedding.csv" is only used to train the model. Then hyper-parameters are adjusted based on the performance of "dev_mbedding.csv" to avoiding fitting. Finally, the real score is marked based on the "test_embedding.csv", which doesn't have ground labels on toxicity. All data sets include one column of ID, one column of label of toxicity, twenty four columns of different identites and 384 columns that represents the word vector of embeddings. Identities selected to split five subsets are "Christian", "Muslim", "Female", "Homosexual gay and lesbian" and "Male". Each subset corresponds to one of the five identites.

Chosen algorithms to classify and compare are Multi-layer perceptron (MLP) and K-nearest neighbors, versus a baseline model, which is simply the "majority vote" (Zero-R).

2. Literature review

In Toxic Comment classification [2], the harm of the toxic comments is well elaborate in terms of "Cyber-bullying", in order to show the important role of toxicity monitoring system on online social platforms. In addition to this, Zaheri et al (2019) [2] also demonstrate and implemented conceptes of Naive Bayes model and Long Short Term Memory (LSTM) / RNN to the data set of WikiMedia Foundation [1]. After processing the data in a specific way, LSTM approach increases almost 20% in True Positive Rate (TRP), comparing with the baseline model Naive Bayes, shown in Table 1.

Comparing with the study [2] , Betty van Aken et al.(2022) [4] implemented more classification models, which include CNN, LSTM and Logistic Regression; sequentially showed that different errors can be made by approaches, however, those can be improved by combining into an ensemble. Moreover, Betty, V. A. et al. explored errors more deeply and manually summarized into

[1] meta.wikimedia.org/wiki/Research:Detox

groups based on the errors of False Negatives and False Positives. As a conclustion, the main reason of most errors is the "lack of consistent qualit of labels" and the missing training data.

Metric	Naive Bayes (NB)	LSTM
TRP	48%	67%
F1	64%	73%
Precision	94%	81%
Recall	48%	66%

Table 1. Summary figure that compares performance of each method

3. Method

In this section, I study models for the classification task in details in the respect of solving problems addressed before. Further, based on the observation, I explore methods to avoid unintended biases. The general goal is to detect toxic comments correctly and equally for different identites. All algorithms and metrics are implemented and generated using python machine learning package called scikit-learn [2].

3.1. Zero-R

The Zero-R model is applied as the baseline to compare the accuracy of predictions. It is simple to implement and intuitively shows the distribution of the data set. In contrast to that, it is skewed towards the majority class. In inbalanced data set, the accuracy can be extremely high.

3.2. K-Nearest Neighbor

In order to properly use the vertor representation of comments, KNN, as geometric classifier, is highly used for tasks of text classification. It doesn't required high level of computing power and meets human intuition, based on the assumption that similar instances are closely located in the space plane. This characteristic is helpful to explore the properties of the data set. To calculate "distance", euclidean distance is implemented. To diminish the effect of "Dimension Curse", the inverse distance is applied as the weight to penalize distant instances.

3.3. Multilayer Perceptrons

MLP, also known as Neural Networks, are recently becomeing more popular in many areas. It is able to distinguish non-linear data and find out potential relationship among datas, which is beyond human understanding. By intuition, it is designed to have 2 hidden layers, and they are in size of 5 and 2 in monotonic decreasing order. The reason behind these size is trying to characterise 5 identities first, then converge the sentence meaning positively or negatively. Since it is a binary classification task, the active function is set to be "logistic" function, which maps output in a range between 0 and 1.

3.4. Evaluation Metrics

Considering the specificity of the toxicity detection, both high rate of wrongly judged to be toxic and low rate of missed detecting toxic comments are not acceptable. The general accuracy gives a good scale on both side, but it can be easily misleaded by inbalanced distribution. Therefore, the recall value is admissible, which measures how many toxic comments the model can detect. Another noticable metric is AUC of ROC, which combines true positive rate and false positive rate to indicate capability of distinguish toxicity over non-toxicity without influnced by data distribution.

3.5. Procedures

To have a general sense of how well the model performance in general, two models are trained by the entire data set and predict on each identity specifically. As a comparison, the training data are selected by the aiming identity, in order to see if the identity has some special characteristics that may be overwhelmed in the entire data set. On purpose of better evaluating results, the chosen K in KNN model and the size of hidden layers of MLP keep constant.

[2] https://scikit-learn.org/stable/

4. Result

In table 2, different perfomance of models for one specific identity is clearly shown. Genearlly speaking, the baseline model gets a very high accuracy but extremely low recall and AUC scores. KNN models get better recall and AUC entirely and Separatly. In contrast, the accuracy is relatively low with maximum 16% difference comparing with the baseline. In the mean time, the MLP algorithm shows it's power as all three metrics are relatively batter than other models, except for Muslim which is higher than KNN though. Especially AUC reaches over 80 percent for Christian, Female and Male.

Nevertheless, the table also shows that there are significant gaps between different identites. Accuracy various from 73 to 90 for KNN and from 77 to 91 for MLP. Indeed, MLP fails to converge when training Separatly within 500 iterations for some identites.

Model	Christian		
	Accuracy	Recall	AUC
Zero-R	90	0	50
KNN (Entirely)	90	21	69
KNN (Separatly)	90	18	69
MLP (Entirely)	91	20	87
MLP (Separatly)*	91	18	86
Model	Muslim		
	Accuracy	Recall	AUC
Zero-R	90	0	50
KNN (Entirely)	74	32	63
KNN (Separatly)	73	30	64
MLP (Entirely)	80	35	80
MLP (Separatly)	78	41	79
Model	Female		
	Accuracy	Recall	AUC
Zero-R	85	0	50
KNN (Entirely)	83	19	64
KNN (Separatly)	83	18	63
MLP (Entirely)	86	22	82
MLP (Separatly)*	87	25	82
Model	Homosexual gay or lesbian		
	Accuracy	Recall	AUC
Zero-R	73	0	50
KNN (Entirely)	71	23	60
KNN (Separatly)	69	27	59
MLP (Entirely)	77	30	73
MLP (Separatly)	76	40	73
Model	Male		
	Accuracy	Recall	AUC
Zero-R	83	0	50
KNN (Entirely)	80	18	62
KNN (Separatly)	81	20	63
MLP (Entirely)	85	25	82
MLP (Separatly)	85	26	81

Table 2. Summary figure that compares performance of each method for identites (*: Stochastic Optimizer can not converge within 500 iterations)

5. Disscussion

Genearlly speaking, all models works poorly to correctly classify toxicity. Majority of toxic comments are missed. The relatively high accuracy is just becasue of the large proportion of non-toxic comments. Even so, there are still some points worth discussing.

5.1. Extreme Results of Zero-R

Considering the methodology of majority voting, it is not surprised to see that the recall is absolutely zero and AUC is at the lowest level of 50 percent, since all instances are labelled as ”non-toxic”. Moreover, the very high accuracies obviously indicates that the distribution of toxic comments is not balanced. It varies from 73:27 to 90:10. It is noticable that inbalanced distribution will bias ”learning model” towards the dominant class, which is non-toxic in this study, such that the capability to detect toxic comments will be negatively influenced.

5.2. Higher Recall vs Higher Accuracy

As one of the goal to achieve, one generalized model that can effectively detect high proportion of toxic comments is expected. However, this is a trade-off between accuracy and the recall, that increasing recall causes accuracy to decrease. This is reasonable since in the inbalanced data set, the model should be more generalized to detect the minor class, meanwhile, the ability that specially target the majority class will lose. This is supported by KNN results that accuracy is relatively lower than the baseline, the one that is extremely targeting on the majority class, but more proportion of toxic comments are detected.

5.3. Undistinguishable Toxic Comments

Thinking of the assumption of KNN and the high accuracy but low recall, it seems that the nearest neighbors are more likely to be non-toxic. In addition to this, the AUC of slightly above 50 percent majority also supports that two classes can not be clearly distinguished. Nonetheless, the majority of non toxic comments are still correctly predicted. Thereby, it is possible to deduce that most toxic comments are located near non-toxic ones, which violate the pre-assumption of KNN method. Furthermore, if the labelled data can be assumed to be confident and the embedding word vectors truly represent the meaning of the sentence, we can say that the majority of toxic comments don't have an obviously feature comparing with non toxic comments. Only seeing the meaning of the sentence is not enough to judge toxicity, since sentences with similar meaning are not toxic! However, actually the data set is not perfectly labelled with high confidence as described above at the beggining. Thus, the lack of certainty of the dataset may be one of the reason to explain poor perfomance for KNN model.

In contrast, MLP doesn't heavily depend on the quantity of features. It tries to find out the potential boundary between classes, which is hard to interpret. Since the recall of this model is generally low as well, it fails to figure out the main difference of toxicity. In other words, toxic comments are hard to distinguish.

5.4. Unintended Bias

As shown in table 2 again, the perfomance of the same model for differenct is not equal. The results of Separatly trained model is intended to show that identites are not equally treated without disturbance of other identites. Accuracy varies from 76 in Homosexual to 91 in Christian for separatly trains MLP, while recall differs from 18 in Christian to 41 in Muslim. Always, it is possible to say that different distribution or different amount of instances is one of the reason that casuse difference. Despite that, various recalls using entire train model tell that Muslim and Homosexual are easier to detect toxicity although Christian and female are least likely detect toxicity. Another evidence is that the recall of Christian and Female does not significantly change training from entirely to separatly, only about 2 to 3 percent. On the other hand, the recall increases 10 percent for Homosexual. In other words, even though the model is generalized for other identites using the entire train set, these two identities are still conspicuous to be considered as toxic.

6. Conclusion

In this study, I explore the different perfomance of each model and trying to explain the reason behind beased on features of the dataset, such as inbalancing and label quality. Overall, the MLP model perfomance best. Furthermore, the bias of chosen identites is discussed and the metrics results indicate that Muslim and Homosexual are more likely to be detected as toxic. However, in order to get more specific answers, further studies should be obtained. For example, manually change the distribution to see if accuracy and recall can be balanced, optimize the hidden layer size of MLP to imporve the overall performance, and so on.

References

[1] Jigsaw/Conversation AI. Jigsaw unintended bias in toxicity classification. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification. Accessed: July, 2022.

[2] Pallam Ravi, Greeshma S Hari Narayana Batta, and Shaik Yaseen. Toxic comment classification. International Journal of Trend in Scientific Research and Development (IJTSRD), 2019.

[3] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

[4] Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. arXiv preprint arXiv:1809.07572, 2018.