Improvement of toxic language classification by using unlabelled data

Anonymous

1. Introduction

Internet could be an ideal place for communication free talking. However, since there are minimal consequences from bad behavior, rudeness, hateful speech, and discrimination etc. toxicity language on internet is a major source of online harassment and cyberbullying. These comments that contains explicit language could hurt (in many ways) others [1].

Online communication channels could use filters to filter out toxic comment and block users who produce them. However, no known technic or model can successfully recognize all toxic comment, the reason being: 1) human language is complex for machines to fully understand, 2) amount to data need to processed is enormous, therefore it is not feasible to verify them manually [2].

This paper aims to find out can unlabelled data help to improve the prediction performance regarding language toxicity.

2. Literature review

2.1. Toxicity classifier performance

In the case of toxicity classifier, a false prediction could do more harm. In this article [7], one user stated his race and sexual preference and the user got blocked by the system, as the system predict his statement is toxic. This kind of false prediction did more harm than letting a few toxic comment slips through.

Therefore, the key evaluation measurement should be precision, instead of accuracy & recall. Precision is a measurement regarding how good the model is at predicting true positive (in this case, toxic) over true positive combined with false positive (the case above). This paper will use precision as the key measurement to guide the research.

2.2. Machine learning methods

2.2.1 Supervised learning

Supervised learning is to use labelled data [x, y] to train a specific model, then use the model to predict unlabelled data [x]. Many well-known training models are designed for supervised learning such as Naïve Bayes and Logistic Regression. In this research [4], the author stated that supervised training could hit a high accuracy of 87.14%.

However, one drawback of using supervised learning is that labelled data is expensive. Data scientists often deal with millions (if not billions) data. Labeling dataset requires manually works. This can be quite costly.

2.2.2 Unsupervised learning

On the other hand, unsupervised learning utilized unlabelled data. Instead of predicting the label of each data instance, unsupervised learning divide dataset into clusters, each cluster contains data instances with similarities identified by training model.

The drawback of unsupervised learning is that the performance (accuracy in general) is not good enough to compare with supervised learning. According to this article [5], in a comparison of unsupervised learning, the overall accuracy of unsupervised learning is at about 61%, far lower than supervised learning.

2.2.3 Semi-supervised learning

Semi-supervised learning utilized both labelled and unlabelled data. So ideally it can combat the drawbacks from above two learning models [6].

The target of using semi-supervised learning to use only a small amount of labelled data and a large amount of unlabelled data to collectively train a model. It iteratively predict the unlabeled dataset, then combine current labelled dataset with high confidence newly labelled data for next iteration.

3. Method

3.1. Datasets

The datasets used in this paper is from the Kaggle competition - unintended bias in toxicity classification [7], pre-processed by the teaching team. The data used in this paper includes one training dataset (140,000 instances with toxicity classifier), one development dataset (15,000 instances with toxicity classifier) and one unlabelled dataset (200,000 instances without toxicity classifier. In labelled dataset, non-toxic comment and toxic comment distribution is around 80: 20.

Based on the observation of the label distribution, one assumption of this research is that in a real-life scenario (refer to section 5.3 for more details), the amount of non-toxic comment and toxic comment is also distributed at 80:20, as seen in diagram 1 below.

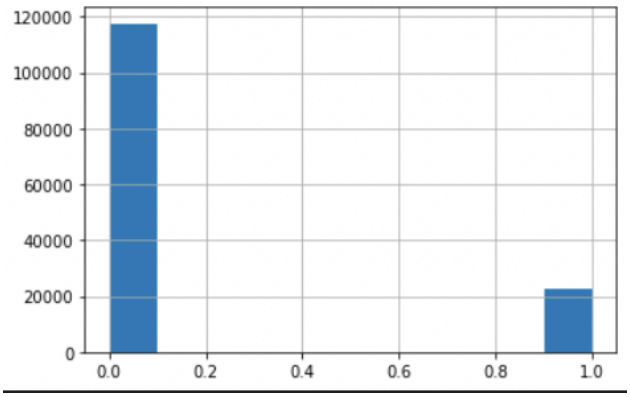

Diagram 1 – Label distribution

Each instance consists of ID, toxicity label, 24 identifiers, 384-dimensional embedding computed comment from original comment.

One instance:
{ID, label, id_1, id_2, … cmt_emb1, cmt_emb2, …}

3.2. General methods

This paper is focusing on researching whether unlabelled data can be used to improve the comment toxicity classifier.

First, this paper set up a baseline mode as the performance baseline.

Second, this paper chooses one supervised learning model, which utilizes labelled data, to train a toxic comment classifier. And chooses one unsupervised learning model, which utilizes unlabelled data, to train a toxic comment classifier. The performance of these two classifiers is compared to find out if unlabelled data can provide better performance.

Then, this paper chooses one semi-supervised learning model, which utilizes both labelled and unlabelled data, to train a classifier. And compare its performance with above two classifiers to find out if unlabelled data can help to improve supervised learning model's performance.

As analyzed in section 2.1, the main evaluation measurement is precision rate in this case. This paper use accuracy as the helper to evaluate the overall performance.

3.3. One-r baseline

This paper use one-r as the baseline model. The dataset provided include a serious of features that is generated based on the original comment. It flags out if the original comment contains a certain word or a word phrase that provide similar meaning. This paper iterates through each identifier to check this identifier's matching-rate to its toxicity label and use the one with highest matching-rate as the baseline to predict the predict dataset.

This method provides a baseline performance. It is expected to observe that other supervised learning model providing higher performance.

3.4. K-mean

This paper use k-mean as the model for unlabelled data under unsupervised learning. Since the toxicity label is binary (either toxic or non-toxic), the cluster number is being set to 2. This paper combines the 140,000 instances of labelled data (with label chopped) and 200,000 instances of unlabelled data, total 340,000 instances of unlabelled data to train k-mean model. Then, this paper uses the trained model to predict 15,000 instances of prediction data.

This method provides the prediction result performance of a fitted unsupervised learning model. This is used to compare with supervised learning model performance.

3.5. Logistic regression (labelled data)

This paper uses logistic regress as the model for supervised training. The model used in this paper is imported from sklearn, with a serious of configurable parameters [8]. First this paper fine

tunes the parameters to get the best performance from given training dataset. This paper iterates through different c-value (regularization strength, where smaller c-value indicate stronger regularization) from [0.001, 0.01, 0.1, 1, 10, 100, 1000]. Then this paper uses the best performed c-value as the parameter for the model, fitting the training data and training label (140,000 instances). Third, this paper use fitted model to predict the result of prediction dataset and compare the result with prediction dataset label (15,000 instances).

This method provides the performance of a classical supervised learning model. This performance will be used to compare with semi-supervised learning model in the following section.

3.6. Logistic regression (with unlabelled data)

To use logistic regression with unlabelled data, this paper uses the self-training classifier imported from sklearn [9]. This classifier uses one conventional supervised learning model as the base model, then it predicts the unlabelled training data (for each instance, a confidence level in use). Based on the confidence level, it then combines the previously fitted labelled data, and newly generated labelled data to form a new model, to predict over left-over unlabelled data. It does this iteration until no new data label can be confidently generated. Then, this paper uses the fitted semi-supervised model to predict the result of prediction dataset and compare the result with prediction dataset label (15,000 instances).

This method provides the performance of a semi-supervised learning model. This performance will be used to compare with above learning models.

4. Result

4.1. One-R baseline

Table 1: One-r baseline performance

Best perform identifier	Precision	Accuracy
"Black"	35.05%	76.36%

From all 24 identifiers provided in the prediction dataset, the identifier "Black" turns to be the best identifier with highest precision rate of 35.05%. This will be our baseline when evaluating other models, refer to table 1 above.

Full result please refer to the program files.

4.2. Supervised learning models

This paper first evaluates three conventional supervised learning method preliminarily: naïve bayes, logistic regression and neural network. The below result (refer to table 2) indicate that logistic regression has the best performance with a precision percentage at 81.43% and a accuracy percentage at 81.39%. This is expected and discussed in 3.6.

Table 2: Supervised learning models performance

	Precision	Accuracy
Naïve bayes	33.60%	68.35%
Logistic Regression	81.43%	81.39%
Neural network	9.0%	60.0%

4.2.1 Logistic regression in details

Table 3: Detailed logistic regression model results

	Precision	Recall	F1-score
Non-toxic	81.39%	99.89%	89.70%
Toxic	81.43%	2.01%	3.92%
Accuracy			81.39%
Macro avg			46.81%
Weighted avg			73.49%

The above result (table 3) indicates that a tuned logistic regression model could hit a overall accuracy of 81.39% with a relatively high precision score of 81.43%. This is vital to this case, as stated section 2.1, the most important measurement should be precision, so the filter will have less accidents of predicting non-toxic comment as toxic.

Compared to Naïve bayes and neural network, logistic regression is considered to be the better model to predict language toxicity. This paper will use logistic regression as the base model in the semi-supervised learning.

4.3. Unsupervised learning

Table 4: Detailed k-mean (unsupervised learning) model results:

	Precision	Recall	F1-score
Non-toxic	77.62%	71.17%	74.25%
Toxic	8.79%	11.92%	10.12%

Accuracy	59.97%
Macro avg	42.19%
Weighted avg	62.13%

The above result (table 4) indicates that using unlabelled data along (unsupervised learning) doesn't provide a decent performance (both accuracy and precision). In fact, its performance is below the baseline model from section 4.1.

This result is matching the discussion of unsupervised learning in section 2.2.2.

4.4. Semi-supervised learning

Table 5: amount of unlabelled data's effectiveness

Unlabelled data amount	Fraction	Precision	Accuracy
200	.1%	81.42%	81.39%
2,000	1%	81.05%	81.37%
20,000	10%	82.53%	81.37%
40,000	20%	76.19%	81.32%
100,000	50%	71.05%	81.31%
200,000	100%	70.13%	81.30%

This paper feed unlabelled data gradually to the base model to check the effectiveness of unlabelled data. From the above table 5, it shows that after introducing unlabelled data to a semi-supervised training model, the performance is gradually decreasing. However, by given just a small fraction of unlabelled data, it could improve the performance as show in diagram 2 below.

Table 6: Detailed results with 20,000 instances:

	Precision	Recall	F1-score
Non-toxic	81.37%	99.91%	89.69%
Toxic	82.54%	1.83%	3.59%
Accuracy			81.37%
Macro avg			46.64%
Weighted avg			73.42%

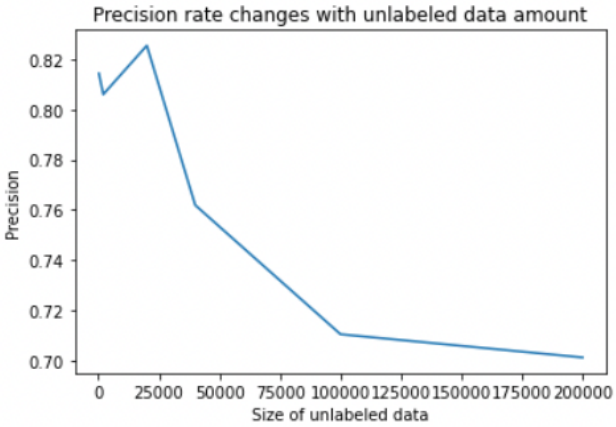

Diagram 2 – semi-supervised learning performance curve

4.5. Summary of results

Table 7: overall performance comparison

	Precision	Accuracy
One-R baseline	35.05%	76.36%
Logistic Regression (supervised learning)	81.43%	81.39%
K-means (unsupervised)	8.79%	59.97%
Logistic Regression (semi-supervised learning)	82.54%	81.37%

In sum, the performance of semi-supervised learning model is quite similar to supervised learning model.

5. Discussion

5.1. Unlabelled data performance

From above results, it can be observed that using unlabelled data along is less likely to achieve a high-performance model in comparison to using labelled data. By using unlabelled data in associate with labelled data (such as semi-supervised learning) with some tuning, it is expected that unlabelled data could improve the performance of predicting language toxicity.

However, the performance gain of using unlabelled data is quite limited and does not provide a convincing result to justify the extra effort spend on using semi-supervised learning is well spent.

5.2. Future improvement for unlabelled data

This research has only conduct experiments of unlabelled data on a few machine learning models. By no means that this research is rejecting the use of unlabelled data. In contrast, this paper does discover that in some scenario unlabelled data could provide improvements on prediction performance of language toxicity.

Therefore, future improvement for unlabelled data could be found in different algorithm and machine learning models.

Furthermore, this research has only conduct experiments in the case of language toxicity. The performance gain using unlabelled data could be much higher in other subjects.

5.3. Limitations

5.3.1 Label balance

One major assumption made in this paper is that in real-life dataset, the distribution of toxic and non-toxic comment label is very similar to the training set label distribution. The model trained in this paper is expected to perform badly if the distribution percentage is far different from the training set.

One way to mitigate this limitation is to adjust the model based on real-life data, since not all online-environment behaves the same. For example, toxic comment is less likely to appear in a research-focused forum than an anonymous gaming chatroom.

5.3.2 Model selection

Naïve Bayes and neural network are only tested preliminarily during model selection stage.

In semi-supervised learning model, only one base mode is tested. Using multiple base models comprehensively could provide better performance.

5.3.3 Data validation

Original data validity is not tested, this paper assumed to trust the data validity.

6. Conclusions

After taking several experiments this paper demonstrates that in the context of language toxicity and given datasets, using unlabelled data along is not as good as using labelled data along. This paper also demonstrates that by using unlabelled data and labelled collectively, the performance could be improved however only by a thin margin in this case.

References

[1] Sara Zaheri, Jeff Leath, and David Stroud. Toxic comment classification. SMU Data Science Review, 3(1):13, 2020.

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

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

[4] V. Nasteski, "An overview of the supervised machine learning methods", www.researchgate.net, 2022. [Online]. Available: https://www.researchgate.net/profile/Vladimir-Nasteski/publication/328146111_An_overview_of_the_supervised_machine_learning_methods/links/5c1025194585157ac1bba147/An-overview-of-the-supervised-machine-learning-methods.pdf. [Accessed: 07- Oct- 2022].

[5] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", International Journal of Advanced Research in Artificial Intelligence, vol. 2, no. 2, p. 34, 2013. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.278.5274&rep=rep1&type=pdf#page=41. [Accessed 7 October 2022].

[6] X. Zhu and A. Goldberg, "Introduction to Semi-Supervised Learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3, no. 1, pp. 1-130, 2009. Available: 10.2200/s00196ed1v01y200906aim006 [Accessed 7 October 2022].

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

[8] "sklearn.linear_model.LogisticRegression", scikit-learn, 2022. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linea

r_model.LogisticRegression.html. [Accessed: 07- Oct- 2022].
[9] "sklearn.semi_supervised.SelfTrainingClassifier", scikit-learn, 2022. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.SelfTrainingClassifier.html. [Accessed: 07- Oct- 2022].