The influence of unlabelled data on Twitter sentiment classification

Anonymous

1 Introduction

In the past 20 years, social media and blogs become increasingly popular. Twitter is one of the biggest blogs where people enjoy sharing opinions and expressing sentiments. Therefore, scientists start to do opinion mining and emotion analysis to find the sentiment of the posts.

It is crucial to train some algorithms to help people label unlabelled data since the number of posts is enormous. There exist algorithms that can be improved. And there are datasets from Twitter, which can be used in the research. In the paper of Blodgett et al. (2016), they make manually annotated Twitter data available. Thus, this paper explores if the unlabelled data can improve Twitter sentiment classification.

The rest of the study includes the following sections: related works, main methods, results, discussions about the research question and a conclusion.

2 Literature review

There exists much paperwork that research on this topic using different methods and approaches.

In Blodgett et al. (2016) research, the manually annotated Twitter data is being used. It is the proper dataset for the experiments because the data are collected from the blogs using the ways of Streaming fashion. So, it is the real tweets that include the language use habits and content from the public.

2.1 Supervised and Unsupervised learning

Supervised Machine learning methods have been implemented in many studies (Agarwal and Mittal, 2016), and in these researches, the fully labelled dataset is used when training the model. Based on Hidayat et al. (2022) paper, the Naïve Bayes, Decision Tree and Random Forest algorithm can take training data and predict unlabelled data which Naïve Bayes gives the highest accuracy of 83.43% (Fitri et al., 2019). However, labelled data are not always achievable, since it is expensive (Huang et al., 2018), and unsupervised learning can be used during this process. But the accuracy of the prediction is still left behind of supervised learning by a large amount (Sazzed and Jayarathna, 2021).

2.2 Semi-supervised learning

To overcome these problems, semi-supervised learning (SSL) can be used for prediction, which uses a small portion of labelled data, and take unlabelled data into the training model. Huang et al. (2018) has claimed that the SSL approach has improved the accuracy by 20%, and eventually SSL achieves an accuracy of 77% (Sazzed and Jayarathna, 2021).

2.3 Summary

From this part of the research, it can be seen that there are datasets available to support the research, and also exists some work that shows how unlabelled data works in supervised, unsupervised and semi-supervised learning. However, none of them includes the proportion of unlabelled data in model training.

3 Method

3.1 Datasets and pre-processing

The data used is from Blodgett et al. (2016) research, and in this dataset, raw data is converted into numbers in different dimensions using the Embedding method. Data preprocessing and wrangling are utilized to prepare training data, testing data and unlabelled data. Data description will make sure if this dataset has balanced data.

3.2 Supervised learning and Unsupervised learning

3.2.1 Baseline

A baseline will be implemented by simply guessing all data have the same label, and the accuracy of any model should be better than the baseline. To show the algorithm is comparatively valuable.

3.2.2 Logistic Regression

A logistic regression (LR) algorithm is a supervised learning model that takes training data (with label) to fit the model. And the classification results will be applied to testing data and predict the label. Theoretically, LR uses Cross-Entropy Loss to measure the distance between two events or distributions and to classify the data.

$$\text{L} = \sum(Y_{true}log(p)) + (1 - Y_{true})log(1 - p)$$

The parameter C (inverse of regulation strength) will be adjusted, to get a comparatively good prediction.

3.2.3 K-Means Clustering

K-means clustering (KML) algorithm is an unsupervised learning model that takes unlabelled data to train the model. And k is 2 here, which represent 'positive' and 'negative' labels. KML algorithm starts with randomly chosen cluster centres, data points will be classified in the nearest centre with some movement.

Two parameters 'n_init' (algorithm run time) and 'random_state' (random centroid initialization) will be adjusted to make the algorithm have relatively better performance.

These two algorithms will provide the predicted labels for the testing data and are ready to be used in the later sections.

3.3 Unlabelled data

Semi-supervised learning is created using the combination of unlabelled data and supervised data, SSL will be generated with a base of LR. To test the proportion of unlabelled data on the model, different numbers of unlabelled data used to train the model and compare the performance of predicted data. A 0.7 threshold will be used on the prediction of unlabelled data, which means only when the confidence level of the predicted result more than 70%, data will be kept and used in the training model. And unlabelled data will treat the same as training data after prediction.

3.4 Evaluation

To evaluate the model, and make the comparison between the results, two commonly standard scores will be used:

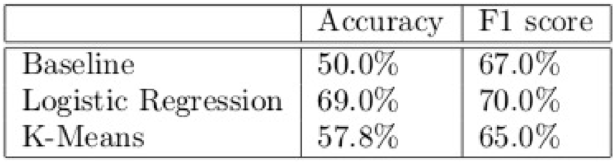
Accuracy: to show how often is the prediction correct using the equation.

F1 score: combines the precision and recall metrics.

Therefore, accuracy and f1 score will value the performance of the models, the higher the better.

4 Result

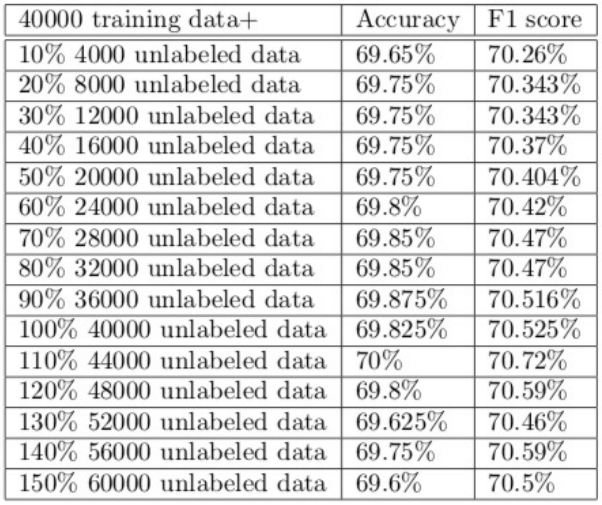
Here is the result get from the code implementation: The accuracy and f1 score for supervised learning LR and unsupervised learning k-means clustering:



	Accuracy	F1 score
Baseline	50.0%	67.0%
Logistic Regression	69.0%	70.0%
K-Means	57.8%	65.0%

Table 1: Accuracy and F1 scores

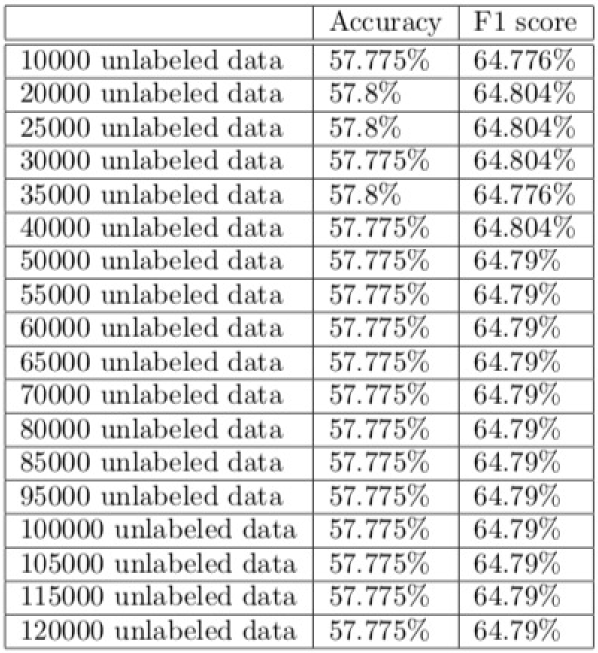
The accuracy and f1 score for a different proportion of unlabelled data for LR:



40000 training data+	Accuracy	F1 score
10% 4000 unlabeled data	69.65%	70.26%
20% 8000 unlabeled data	69.75%	70.343%
30% 12000 unlabeled data	69.75%	70.343%
40% 16000 unlabeled data	69.75%	70.37%
50% 20000 unlabeled data	69.75%	70.404%
60% 24000 unlabeled data	69.8%	70.42%
70% 28000 unlabeled data	69.85%	70.47%
80% 32000 unlabeled data	69.85%	70.47%
90% 36000 unlabeled data	69.875%	70.516%
100% 40000 unlabeled data	69.825%	70.525%
110% 44000 unlabeled data	70%	70.72%
120% 48000 unlabeled data	69.8%	70.59%
130% 52000 unlabeled data	69.625%	70.46%
140% 56000 unlabeled data	69.75%	70.59%
150% 60000 unlabeled data	69.6%	70.5%

Table 2: Accuracy and F1 scores of Logistic Regression

The accuracy and f1 score for a different proportion of unlabelled data for k-means clustering:



	Accuracy	F1 score
10000 unlabeled data	57.775%	64.776%
20000 unlabeled data	57.8%	64.804%
25000 unlabeled data	57.8%	64.804%
30000 unlabeled data	57.775%	64.804%
35000 unlabeled data	57.8%	64.776%
40000 unlabeled data	57.775%	64.804%
50000 unlabeled data	57.775%	64.79%
55000 unlabeled data	57.775%	64.79%
60000 unlabeled data	57.775%	64.79%
65000 unlabeled data	57.775%	64.79%
70000 unlabeled data	57.775%	64.79%
80000 unlabeled data	57.775%	64.79%
85000 unlabeled data	57.775%	64.79%
95000 unlabeled data	57.775%	64.79%
100000 unlabeled data	57.775%	64.79%
105000 unlabeled data	57.775%	64.79%
115000 unlabeled data	57.775%	64.79%
120000 unlabeled data	57.775%	64.79%

Table 3: Accuracy and F1 scores of K-means Clustering

5 Discussion

Based on the research question: if the unlabelled data can improve Twitter sentiment classification. The algorithms have been implemented, and the result is shown in the previous section. Through these results, some points are worth discussing:

5.1 Labelled and unlabelled

In this experiment, from the result section (see Table 1), it is found that the accuracy of supervised learning is 12% higher than unsupervised learning. It is reasonable that, supervised learning uses correct answers (labelled data) to train, and unsupervised learning just clusters all data, and label them without any standard or correct answer. Therefore, when comparing the LR model and K-Means clustering model, LR has better performance. So, labelled data in machine learning is more valuable than unlabelled data and usually provides a better answer.

5.2 Semi-Supervised learning

Based on Table 2, SSL has LR base that provides better performance than supervised learning LR. This can be illustrated by when doing this experiment using SSL, more data is used to train the model. And the additional data has reliability, the predicted answer has a confidence level larger than 70%. Using more reliable data can improve the performance of the model. So, when joining some unlabelled data into a supervised learning algorithm, the accuracy of the LR model has improved.

However, is there any certain proportion of training data and unlabelled data when fitting the model?

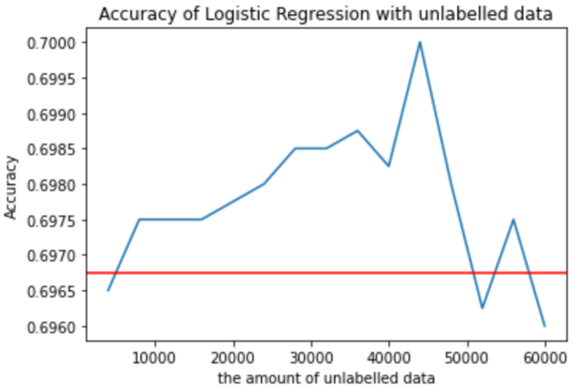


Figure1: Accuracy of Logistic Regression with unlabelled data

From figure above, it can be seen that there is a trend that the accuracy increases with more unlabelled data get involved. When unlabelled data around 40000, the accuracy reaches its maximum. And from the whole proportion, when the unlabelled data between 24000 and 48000, it makes a comparatively good model. So, based on the proportion, train data is 40000, and the unlabelled data between 60% and 120% of training data maximize the performance of the model. But for the unlabelled data in supervised learning, why is not the more the better?

Although after predicting the unlabelled data, it becomes relatively reliable data, the confidence level is 70% still lower than the correct answer. Therefore, as the proportion of unlabelled data increases, the confidence level for the whole training set in SSL decreases. And the potential risk of passing the wrong label or error from the training set to the prediction is still there.

To do a critical analysis of the SSL, there are some drawbacks and limitations of the self-training model. We implemented self-training with a 70% confidence level, therefore during the first round of prediction (label the unlabelled data) there are some of the data that will not be kept as training data. Therefore, the actual proportion used might be smaller than the record to some extent. And the percentage of unlabelled data calculated might be over-valued. Therefore, to improve the experiment we can use other algorithms or functions to record the actual data that is kept under the confidence level.

5.3 Unsupervised learning

From Table 1 in the result section, it can be seen that more datasets do not always give better accuracy.

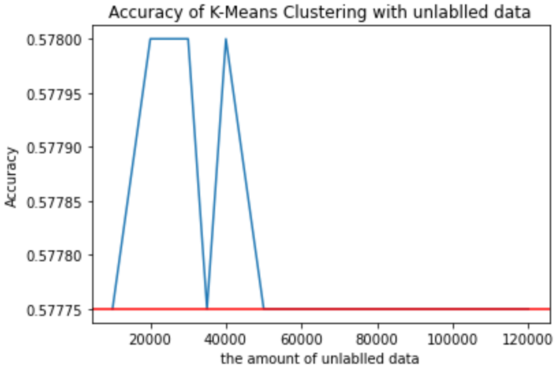


Figure 2: Accuracy of K-Means Clustering with unlabelled data

And figure above suggests that, when using unlabelled data to train the model, the amount of data between 20000 to 40000 will fit the best model. This has happened because too much unlabelled data may have the risk to include more noise data that will lead the model in the wrong direction. And when the unlabelled data is between 20000 to 40000, it can fit the best model in this experiment. This can be illustrated as when the dataset is in this range, it can focus more on the data and correctly classify them into different clusters.

5.4 Limitation of the experiment

Although the experiment is done fully prepared, there are still some limitations when evaluating the performance of the algorithms and unlabelled data.

First of all, when discussing the accuracy, we use high, low accuracy or if unlabelled data can improve the model. However, from the tables above, the actual differences between the accuracy of each model is really small. For example, in this experiment, although semi-supervised learning that involves unlabelled data can improve the performance of the data, the improvement is still less than 1%. That is because, when implementing the model, it uses the same base algorithm and same training data. And also, when using the testing data, the test data always be the same, but the same testing data has unicity. Therefore, even if we keep testing on different models, the accuracy cannot be improved a large amount. So same training data, base algorithms and testing data that makes the accuracy of different experiments do not have that many differences.

Moreover, the testing set used in this experiment has randomness. Thus, all evaluation that currently has is all base on this dataset, if changed to another dataset, the result might be slightly different. Therefore, if the experiment can test on different datasets, the result from this question will be more valuable.

5.5 Summary

Although there are many limitations in the experiment and also many points that can be improved in this research, the result that we get is still valuable and can be summarised in the next section.

6 Conclusions

In this study, both supervised and unsupervised learning are implemented, to evaluate the performance of the algorithm and test the influence of unlabelled data on the classification. The logistic Regression algorithm is a good model to predict Twitter sentiment which gives an accuracy of 69.675% and an f1 score of 69.127%. Using unlabelled data involves in the training, the accuracy is improved by 0.325%, and when unlabelled data is between 60% to 120% of the labelled training data, has fit the best model. For unlabelled data, when there are 20000 to 40000 training data it gets the best training model.

Having carried out the previous points it can conclude that unlabelled data can improve Twitter sentiment classification to some extent.

Some future work that can be done, for example: getting more training data and testing data to make the dataset diverse or using cross-validation. And optimize the algorithms to get the proper proportion of unlabelled data.

References

Agarwal, B. and Mittal, N. (2015). Machine learning approach for sentiment analysis. *Socio-Affective Computing*, page 21–45.

Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Fitri, V. A., Andreswari, R., and Hasibuan, M. A. (2020). Sentiment analysis of social media twitter with case of anti-lgbt campaign in indonesia using naïve bayes, decision tree, and random forest algorithm. *Procedia Computer Science*, 161:765–772.

Hidayat, T. H. J., Ruldeviyani, Y., Aditama, A. R., Madya, G. R., Nugraha, A. W., and Adisaputra, M. W. (2022). Sentiment analysis of twitter data related to rinca island development using doc2vec and svm and logistic regression as classifier. *Procedia Computer Science*, 197:660–667.

Huang, K., Zhou, D., Cong, Y., Xu, W., Wang, W., and Que, Z. (2018). Tensor discriminant analysis with partial label. *Procedia Computer Science*, 131:416–424.

Sazzed, S. and Jayarathna, S. (2021). Ssentia: A self-supervised sentiment analyzer for classification from unlabeled data. *Machine Learning with Applications*, 4.