Assignment 3: Sentiment Classification in Tweets for COMP90049 Report

Anonymous

1 Introduction

This is a report of the Comp90049 assignment, to introduce the method and observations that I discovered do analysis the sentiment of tweets.

The data include three types, raw data, TFIDF and Embedding. We will find the performances of these data on different model.

We will use different model to train the data, include Gaussian Naïve Bayes, K-NearestNeighbor and Logistic Regression, and mix the methods, then compare to the 1-R DecisionTreeClassifier as the baseline to examine their performance. We will also process the date, like clean the data and feature engineering.

We will talk about the research question 1. To find out if the unlabeled can increase the performance.

2 Literature review

Artical "Using clustering analysis to improve semi-supervised classification" cited that semi-supervised learning is a kind of learning using a small number of labeled examples and a large number of unlabeled examples. Some learning problems is not suitable to use supervised and unsupervised algorithms. They cannot run effectively. So, semi-supervised can play an important role to run. As such, this is a learning problem between supervised and unsupervised learning.

It is suitable when labelling is hard or expensive. The symbol of a suitable semi-supervised learning algorithm is that it has better performance than a supervised learning algorithm. This book uses The Reuters 21578 Distribution 1.0 data set.

Artical "Semi-supervised learning" points that the core idea of the semi-supervision is using different strategy to treat different data. For the labeled data, the algorithm is the same as the algorithm for the tradition supervised learning, and for the unlabeled data, the algorithm will minimize the difference in predictions between other similar training examples.

Supervised training updates model weights to minimize the average difference between predictions and labels. Unsupervised learning, on the other hand, will cluster the similar points together. However, without labels guide, the accuracy is a not acceptable. With insufficient data or hard clustering, both of these two methods will fail to get successful results.

In semi-supervised learning, labeled and unlabeled data can be used together to get a better result. Labeled data play a role as a sanity check; they add structure by establishing the number of classed, and cluster the predictions correspond to right class.

Unlabeled data provide sufficient context. We can estimate the shape of the total distribution by exposing out model through as much data as possible.

As a consequent, we can train more accurate and reliable models.

This paper use Moons dataset example as the dataset. To compare the results by using supervised, unsupervised and semi-supervised learning algorithm.

3 Method

As mentioned before, we have three types of data, raw data, TFIDF and embedded data. And all of these three kinds of data include four parts, train, dev, test and unlabeled. The data comes from "Demographic Dialectal Variation in Social Media: A Case Study of African-American English".

3.1 Feature Engineering

3.1.1 Clean data

Our full data have the text attributes and it contains many tweets. The context of the tweets includes many unnecessary information. For example, "_TWITTER-ENTITY_ i go at 4 lol" includes '_TWITTER-ENTITY_' and this part could be deleted. Also, there are also some other parts can be deleted, such as "#", "https".

3.1.2 Feature added

We will use sentiment from TextBlob for the full data to get two more features, subjectivity and polarity. For GaussianNB and Logistic Regression, the accuracy increases from 61.475% and 69.825% to 61.65% and 69.95% separately. But for KNN, the accuracy decreases from 69.15% to 68.875%.

3.2 Machine learning models

3.2.1 Baseline model - One-R Decision tree

We choose the One-R decision tree model as the baseline. We use the train embedded as the train, and use development embedded as the test. The accuracy is 59.925%. The reason we choose this model as the baseline is that it is easy to implement and comprehend. And it normally has a good result.

Gaussian Naive Bayes is suitable for features that are normally distributed. It assumes the features have strong independence assumptions and the value of specific feature is independent of the value of other features. Especially in supervised learning situation, Naïve Bayes Classifiers are trained very efficiently. Naïve Bayes Classifiers only need a small number of training data to estimate the parameters which are needed for classification. In addition, Naïve Bayes Classifiers can be designed and implemented simply, and they are suitable in many situations.

Using the same data as the baseline, the accuracy is 61.475%

3.2.2 K-NearestNeighbor model

K-NearestNeighbor(KNN) model is the second model we chose. KNN is a simple supervised machine learning algorithm, which can solve classification and regression problems.

To implement the KNN method, it is important to use a suitable K value. I use a for loop to calculate the accuracy of KNN when using different number of K on train embedded data; and K = 46 have the highest accuracy.

KNN is quite easy to implement and use and easy to understand. However, the calculation is relatively large. Because for each text to be classified, the distance from it to all known samples need to be calculated, and its K nearest neighbors can be obtained. And the choice of K is not the same for every learning.

3.2.3 Logistic Regression model

Logistic Regression is the third model we choose. This is another useful supervised machine learning algorithm for binary classification problems. Logistic regression uses linear regression for classification problems.

Logistic regression uses a logistic function to model a binary output variable. To compare with the linear regression, the mainly difference is that logistic regression's range is bounded between 0 and 1. Also, logistic regression does not need a linear relationship between inputs and output variables.

To compare with two other model, logistic regression has a relatively higher accuracy, 69.825

Figure 1 and Table 1 shows the accuracy of baseline and three models' accuracy when using train embedded as training data and using development embedded as test data.

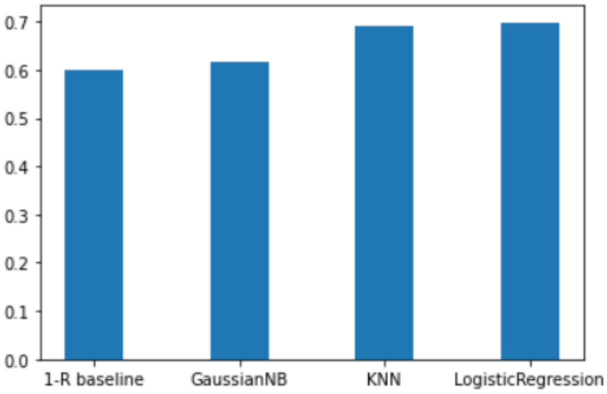

Figure 1: Accuracy of baseline and 3 models

Model	Accuracy
1-R baseline	59.925%
Gaussian NB	61.475%
KNN	69.15%
Logistic Regression	69.825%

Table 1: Accuracy of baseline and 3 models

3.2.4 Combination of models

We will use a combination of GaussianNB, KNN and Logistic Regression. And this combination model has a more accurate result. We will discuss more specific in the next section

4 Result

In this part, we will discuss some result when using different models on different data.

4.1 Add two new features

As mentioned before, we add two new feature, subjectivity and polarity. As shown on Figure 2, the accuracy on Logistic Regression and Gaussian increase. As a consequence, we should not add subjectivity and polarity to test data when using KNN algorithm.

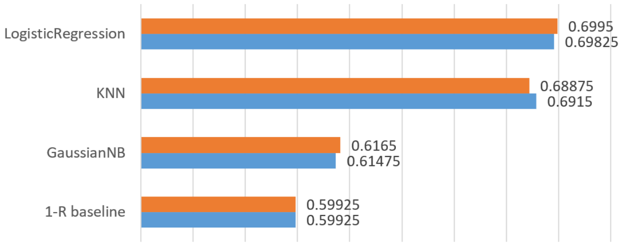

Figure 2: Accuracy after adding two new features

4.2 Select N best feature

We will use two methods, chi and mutual_into_classif, to select N best feature. First of all, we need to use MinMaxScaler to change all of the data to positive. Then, use a for loop to find the accuracy performance when choose 100, 200, 300 and all of the data. By using GaussianNB, the accuracy is shown on Figure 3. It is obvious that when we choose 100 by using mutual_into_classif, the accuracy is the highest, achieved 62.78%.

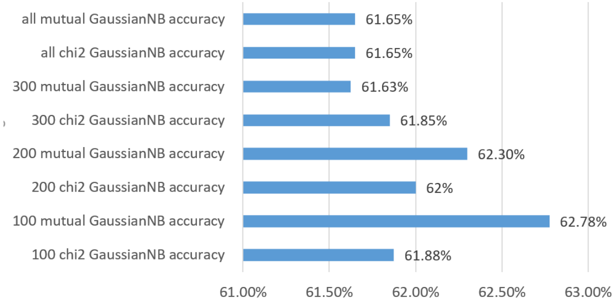

Figure 3: Accuracy for Gaussian

We also use KNN to calculate accuracy after select N features and the result is shown on Figure 4. When we select 300 by using mutual_into_classif, the accuracy is the highest, 69.15

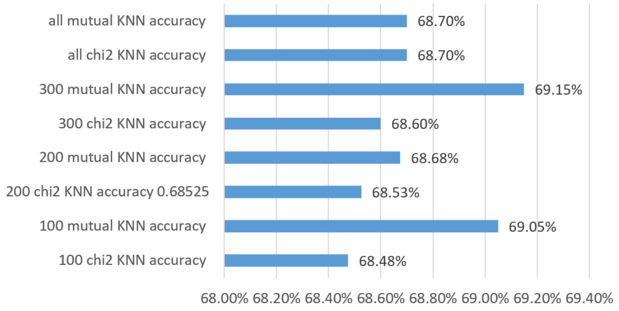

Figure 4: Accuracy for KNN

In the end, we use Logistic Regression. And the result is shown on Figure 5. The highest accuracy is using all of the data.

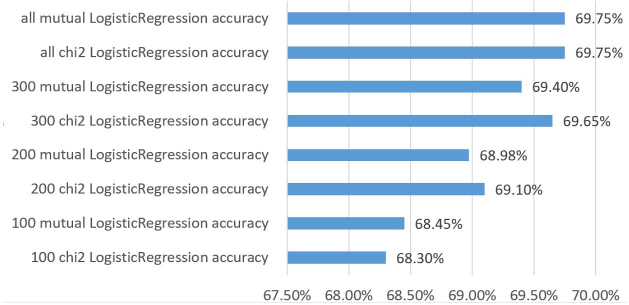

Figure 5: Accuracy for KNN

However, the highest accuracy is smaller than the accuracy without selects N best feature. The reason is that chi2 and mutual_into_classif is suitable for classification algorithm. Both KNN and Gaussian NB are classification algorithm. Logistic Regression is regression algorithm. As a result, it is not suitable for Logistic Regression. Also, we can notice that in Gaussian NB and KNN, the mutual have a better performance than chi2 method.

4.3 Stacking Classifier

Stacking is a technique for ensemble learning, which for combining multiple classification models by meta-classifier. First of all, every classification model will be trained individually. And then, the meta-classifier is fitted based on the outputs. The meta-classifier can be trained on the predicted class labels as well as probabilities from the ensemble.

When we use m1 = GaussinaNB, m2 = KNN and m3 = Logistic Regression, use the train embedded data train model and use development embedded to test, the accuracy is 70.9%, which is higher than any single model accuracy.

As a result, we will choose Stacking Classifier as the model to predict the test. We select 100 features by mutual for GuassianNB as m1, 300 features by mutual for KNN as m2 and Logistic Regression as m3. The result have the highest accuracy in this case.

5 Critical Analysis

We will discuss "Does Unlabeled data improve Twitter sentiment classification?"

First of all, we have some symbol for semi-supervised classification. L is the set of labelled training instances. U is the set of unlabeled training instances. Most of the time, L is larger than U and we need to have a better performance for using L and U together than from L alone.

We need to know if it is necessary to use unlabeled data firstly. Before this, we need to know what is the advantage of semi-supervised learning. Most of time, it is easy to collect data, but it is difficult to label it. Labeling data will cost many times and effort.

To implement semi-supervised training. An important step is self-training. We need allocate all of the labeled and unlabeled data. At the beginning, only use the labeled data to train the model. And we use this model to predict the unlabeled data. For the result which predict.prob(some of the model have this parameter, for example,Logistic Regression,SVC,Random Forest) is positive, we can move them to the labeled data. And then, we do it again, use the labeled data and positive unlabeled data to train the model, then predict. In this step, we need to set a value for the threshold. As shown in Figure 6, the threshold in 0.7 and 0.8 have a vary close result. But KNN and GaussianNB does not have acceptable result to compare with only labeled training results.

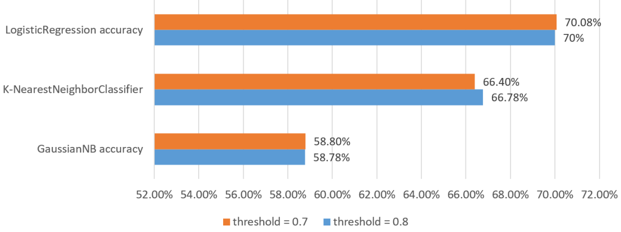

Figure 6: semi-supervised accuracy with different threshold

Based on this result on Figure 7, we can make a assumption, regression algorithm have a better performance than classfication algorithem. So, when we choose to use regression algorithm, we can use some unlabeled data by self-training to improve the accuracy. However, when we use classification algorithm, the unlabeled data will cause negative influence.

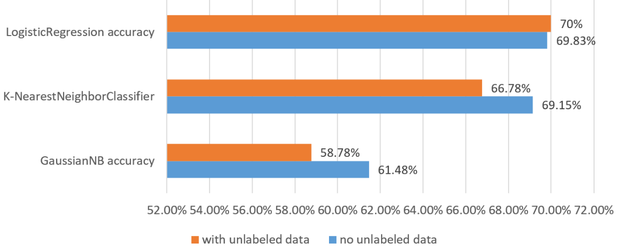

Figure 7: accuracy with labeled and unlabeled data

We still something needs to be noticed. We only use one type of data to do the accuracy test and it may not accurate enough to conclude the performance of different model in semi-supervised learning. To improve the test, we should use two more kinds of data and use the same model to get an average.

6 Conclusion

In this assignment, we use the dataset in different types to analysis the sentiment of data. We use 1-R decision tree as baseline and compare with Gaussian NB, KNN and Logistic Regression. We do some feature engineering and use semi-supervised learning (use labeled data and unlabeled data together) to improve the accuracy. And finally, we use the stacking classifier to get the highest accuracy. In the process of test, we find out that to do semi-supervised learning, accuracy of regression algorithm (include Logistic Regression) improved. But accuracy of classification algorithm (include Gaussian NB and KNN) decreased after using semi-supervised learning.

Concluding text.

References

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.

Chapelle, O., Scholkopf, B., and Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Gan, H., Sang, N., Huang, R., Tong, X., and Dan, Z. (2013). Using clustering analysis to improve semi-supervised classification. *Neurocomputing*, 101:290–298.