NLE

ASSIGNMENT 2

ABSTRACT

Creating a text classifier using SVM, Random Forest and logistic Regression on IMDB movie sentiment dataset. The process was to read the file, find which preprocessing is suitable, clean the data. This assignment report highlights the multiple preprocessing techniques and model selection used through trial-and-error method.

Riffat Siddiquie 2201532

CE 314/887

Abstract

Creating a text classifier using SVM, logistic Regression and Random Forest on IMDB movie sentiment dataset. The process was to read the file, find which preprocessing is suitable, clean the data. Then use the preprocessed data to train the model into predicting if the review is negative or positive.

This assignment report highlights the multiple preprocessing techniques and model selection used through trial-and-error method. SVM, Random Forest and Logistic Regression were used to find the accuracy and results of model built.

Introduction

A movie review is an article reflecting its writers' opinion about the movie and leaving a constructive or destructive opinion, which allows to understand the overall concept of a movie and provide a better recommendation (M. Yasen and S. Tedmori)

Sentiment analysis is one of the tricky tasks of NLE. To be able to make a model understand the difference between negative and positive reviews, the model needs to learn the difference. After training must be tested to evaluate the performance of the model trained. This work has been done using SVM, Random Forest and logistic Regression to compare the working of the models.

Data Preprocessing

The data contained 50% positive and 50% negative examples which was unbiased. Biased data preprocessing was not required, and the data was balanced.

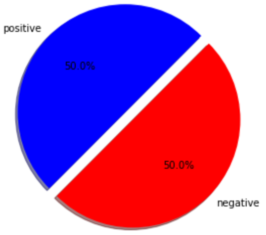
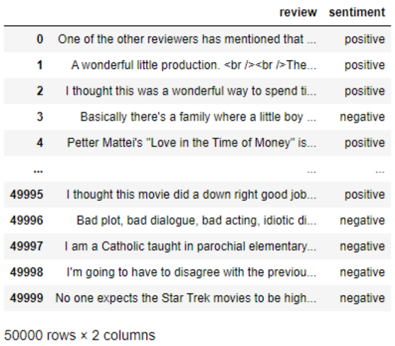


Fig 1: IMDB data distribution

Converting the sentiments to integer as the processing for classification would be lot faster. The positive examples are mapped as 1 and negative as 0.



	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. <br /><br />The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
49995	I thought this movie did a down right good job...	positive
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative
49997	I am a Catholic taught in parochial elementary...	negative
49998	I'm going to have to disagree with the previou...	negative
49999	No one expects the Star Trek movies to be high...	negative

50000 rows × 2 columns

Figure 2: Raw data

Figure 2 shows the raw form which have lots of noise and hard to process data.

Figure 3 shows the steps taken to build a text classification model for sentiment analysis.

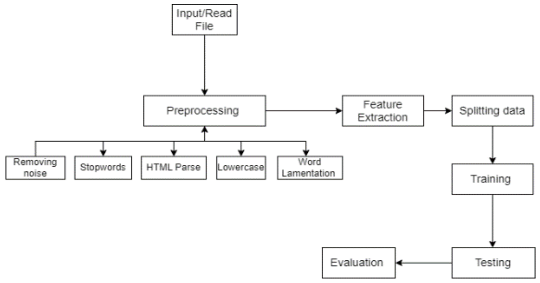
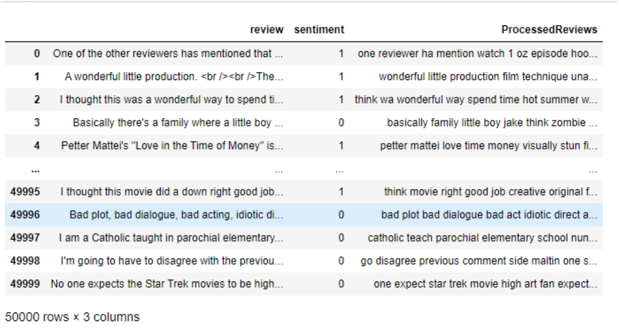


Fig 3: Sentiment analysis procedure

The preprocessing had the steps such as removing the noise (removing special characters, digits), stopwords such as "has", “is” etc. Removing any HTML tags and characters. Converting all the words to lowercase and lemmatization to make sure different spelling but of same origin are categorized as same words.



	review	sentiment	ProcessedReviews
0	One of the other reviewers has mentioned that ...	1	one reviewer ha mention watch 1 oz episode hoo...
1	A wonderful little production. <br /><br />The...	1	wonderful little production film technique una...
2	I thought this was a wonderful way to spend ti...	1	think wa wonderful way spend time hot summer w...
3	Basically there's a family where a little boy ...	0	basically family little boy jake think zombie ...
4	Petter Mattei's "Love in the Time of Money" is...	1	petter mattei love time money visually stun fi...
...	...	...	...
49995	I thought this movie did a down right good job...	1	think movie right good job creative original f...
49996	Bad plot, bad dialogue, bad acting, idiotic di...	0	bad plot bad dialogue bad act idiotic direct a...
49997	I am a Catholic taught in parochial elementary...	0	catholic teach parochial elementary school nun...
49998	I'm going to have to disagree with the previou...	0	go disagree previous comment side maltin one s...
49999	No one expects the Star Trek movies to be high...	0	one expect star trek movie high art fan expect...

50000 rows × 3 columns

Fig 4: Preprocessed data

After the data was cleaned, it has been split into training and testing with ratio of 80%-20%. The first 40000 instances in training and last 10000 as testing. With the shuffle assigned as false to have the same instances each time. (E. Park, J. Kang, D. Choi, and J. Han)

Count Vector is used to find the times the word appeared in the dataset. Using this the most used word in the dataset becomes a feature. The values are further processed in Unicode to reduce computational efforts while training the algorithm. (V. Kumar and B. Subba)

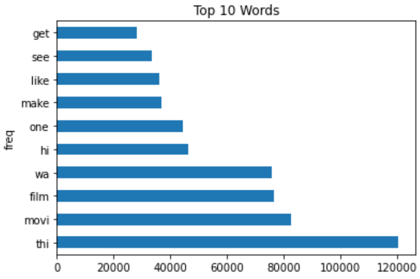


Fig 5: Features of Dataset

Sentiment Analysis

The data has enough examples to show the model which reviews can be categorized as positive and negative. SVM, Random Forest and Logistic Regression both have been implemented to evaluate model’s performance.

SVM treats the instances and nodes and separates them according to their labels. The decision boundary shows the data can be classified.

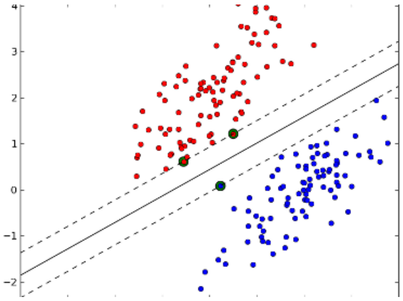


Fig 6: SVM classification

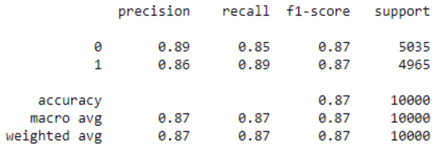
Source: https://ranithsachin.files.wordpress.com/2018/05/svm-451x270.pngused

The test data can be mapped into the trained examples and classified as positive or negative reviews.

The logistic regression model predicts based on the previous trained example that which test node is likely to be in which class, negative or positive.

Results

The results of test of SVM model are the following



```
              precision    recall  f1-score   support

           0       0.89      0.85      0.87      5035
           1       0.86      0.89      0.87      4965

    accuracy                           0.87     10000
   macro avg       0.87      0.87      0.87     10000
weighted avg       0.87      0.87      0.87     10000
```

Fig 8: SVM classification report

The overall accuracy of the model is 87% with precision of 89% and recall of 85%.

The confusion matrix shows the

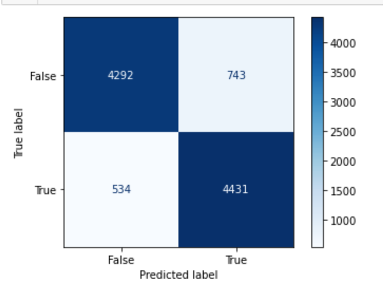
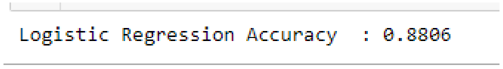


Fig 9: confusion matrix of SVM

The FP and TP both major portion in classification. This is resulting in high performance of the algorithm.

The following are the results are of Logistic regression model evaluation



```
Logistic Regression Accuracy  : 0.8806
```

Fig 10: Accuracy of Logistic Regression

The accuracy of Logistic Regression is slightly better than SVM. The confusion matrix also has higher recall in Logistic regression than of that of SVM

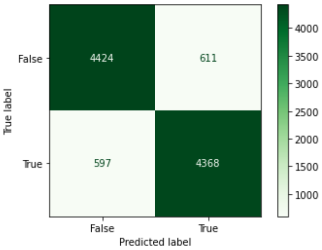
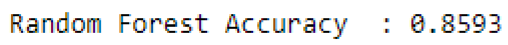


Fig 11: confusion matrix of Logistic Regression

The overall FP, TN and TP are higher than that of SVM.

The following figure shows the results of Random Forest Classifier



```
Random Forest Accuracy  : 0.8593
```

Fig 12: Radom Forest accuracy score

The figure 13 shows the confusion matrix for Random Forest

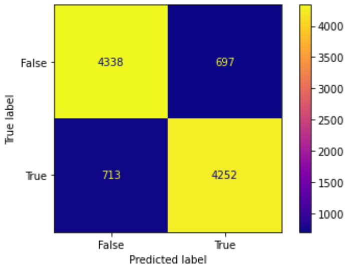


Fig 13: Random Forest Confusion Matrix

Conclusion

Sentiment analysis on IMDB dataset which have labeled examples is preprocessed, cleaned, features are extracted and then the models are fitted for training. SVM, Random Forest and Logistic Regression showed promising results. However Logistic Regression exceeds both other models with accuracy of 88%

Reference

M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 860-865, doi: 10.1109/JEEIT.2019.8717422.

M. Yasen and S. Tedmori, "Movies Reviews Sentiment Analysis and Classification," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 860-865, doi: 10.1109/JEEIT.2019.8717422.

V. Kumar and B. Subba, "A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus," 2020 National Conference on Communications (NCC), 2020, pp. 1-6, doi: 10.1109/NCC48643.2020.9056085.

R. R. Subramanian, N. Akshith, G. N. Murthy, M. Vikas, S. Amara and K. Balaji, "A Survey on Sentiment Analysis," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 70-75, doi: 10.1109/Confluence51648.2021.9377136.

E. Park, J. Kang, D. Choi, and J. Han, “Understanding Customers' Hotel Revisiting Behaviour: a sentiment analysis of Online Feedback Reviews,” Current Issues in Tourism, vol. 23, pp. 605-611, 2020, doi: 10.1080/13683500.2018.1549025.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts Stanford University Stanford, CA 94305

S. Tripathi, R. Mehrotra, V. Bansal and S. Upadhyay, "Analyzing Sentiment using IMDb Dataset," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), 2020, pp. 30-33, doi: 10.1109/CICN49253.2020.9242570.