CE314/887 Assignment 2 Report

Text Classification

Submitted By:

- Muhammad Osama Khan: MK22327

Submission Date: 9th January, 2023

Abstract

A different network for text classification is proposed in this paper. The proposed models have unique and distinct traits, which include the following: (i) tokenize sentence into words (ii) TFIDF, enabling it to attend differentially to more and less important content when constructing the review representation. The experimental approach is taken into account on one large scale text classification tasks which is further evaluated using tfid and tokenization technique that's why model performs well. The results reflect clearly in confusion matrix

1 Introduction

Text classification is an essential part of Natural Language Processing. In the era of digital technology, it is pertinent to properly categorize and classify text in order to gain useful information. By applying the process of NLP on texts, it can convert the text by first analysing it and then implementing labels against the text groups assigned to the relevant text category. It groups and categorizes data efficiently and in a fast manner as compared to manual procedures of handling data. There are multiple approaches which include multiple models, deep-learning process, application using linear models and regression models. [1]

We define a new design, i.e., Logistic Network, which intakes review structure. The design is based on a hierarchal model from word forming sentences. Additionally, the review representation is converted and transformed from basis of counting null values.

2 Logistic Regression

One of the machine learning algorithm methods has a model application termed as Logistic Regression. [2] It is a statistical method which consists of dependant variables as well as independent variables known as predictors. It can be combined with classification to train, test and evaluate the classifier using multiple approaches such as accuracy, confusion matrix, precision, recall. In our paper, we evaluate using the sklearn metrics library

3 Experiments

3.1 Dataset

The imdb dataset is loaded with the relevant libraries imported. The data is analysed using exploratory data analysis to visualize the details of dataset. Furthermore, boxplot and word cloud are implemented to explore the points of dataset.

The dataset is total of 5000 rows further split into train test split feature, with 10% percent as the test data and 90% train. The data is pre-processed using multiple techniques such as lower-casing, removing html, removing brackets stop words. Three models are applied i.e., logistic regression, random forest and stochastic gradient crescent.

Our dataset is interpreted for its effectiveness of model implemented on it. It is done through two document classification datasets. The datasets can be sub followed into two categories:

- Review classification

- Sentiment Estimation

Table 1 reflects the statistical data review in which test train split is 90 % for training and 10% for testing purpose

IMDB reviews are obtained from (Diao et al., 2014). The review ranges from 0 to 1 (negative or positive)

3.1 Baseline

We compare Logistic regression with stochastic gradient descent and random forest The baseline results are reported

3.1.1 Logistic Regression

Logistic regression uses the classify document using the feature

TFIDF used the most frequent 6677306 n-grams (up to 3 grams)

3.1.2 Stochastic gradient descent

Stochastic gradient is optimization method with suitable smoothness it replaces actual gradient calculated by entire dataset

3.1.3 Random Forest

Random forest classifier on criterion entropy and estimators are 50

TFIDF dataset is used on random forest model

Dataset	Classes	Count	Mean	Std	Min	25%	50%	75%	Max
Imdb sentiment	2	50000.000000	0.500000	0.500005	0.000000	0.000000	0.500000	1.000000	1.000000

3.2 Model configuration and training

We split sentence into tokens. We only retain words with help of word cloud visualization Then tfidf vectorizer is used to frequent 6677306 n-grams (upto 3 grams)

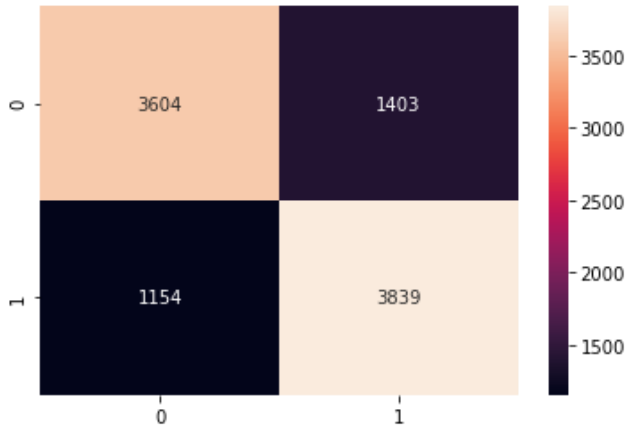
3.3 Result and analysis

	Methods	IMDB DATASET ACCURACY
Zhang et al, 2015	Majority	17.9
	SVM + unigram	39.9
	SVM + bigram	40.9
	SVM + Textfeatures	40.5
	SVM + AverageSG	31.9
	SVM + SSWE	26.2
Tang et al 2015	Paragraph wector	34.1
	CNN word	37.6
	Conv GRNN	42.5
	LSTM GRNN	45.3
This paper	Logistic Regression	74.43
	Stochastic Gradient descent	51.11
	Random forest	51.11

3.5 Visualization

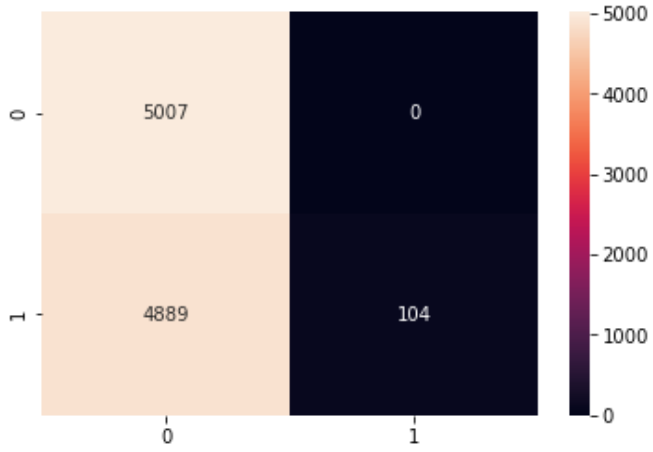
In order to validate our model, we use confusion matrix to test our model.

Logistic regression:



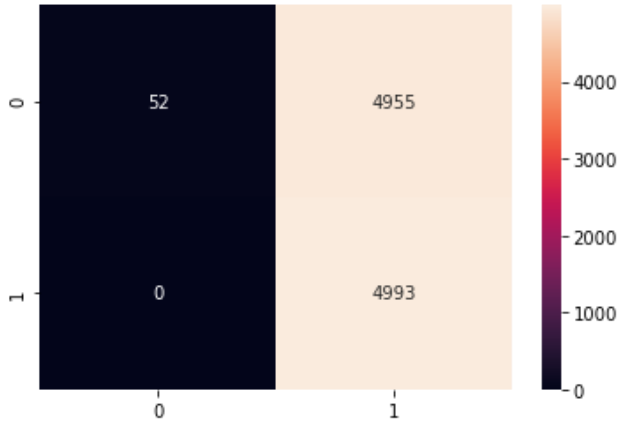
Actual data and predicted by model correctly in 3604 and 3839 whereas 1154 and 1403 our model predicted wrong compared with actual data.

stochastic gradient descent:



Actual data and predicted by model correctly in 5007 and 104 whereas 4889 our model predicted wrong compared with actual data.

Random forest:



Actual data and predicted by model correctly in 52 and 4993 whereas 4955 our model predicted wrong compared with actual data.

4 Related work

Unlike word level modelings, Zhang et al. (2015) apply a character-level CNN for text classification and achieve competitive results. Tai et al. (2015)

5 Conclusion

To conclude, the data provided was visualized and three models were implemented on it. The models included linear regression, random forest and stochastic gradient descent. The results of the models were evaluated using confusion matrix to validate the model's application*. The models showed satisfactory results, therefore, proving it to be a better application of model than previous models.*

References

[1] https://www.sciencedirect.com/topics/computer-science/text-classification

[2] https://link.springer.com/article/10.1007/s41133-020-00032-0