Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.03
Pages: 145-152
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.3.004

Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)

Ayushi Mitra
College of Engineering and Technology, Bhubaneswar,
Bhubaneswar, Odisha, India.
Email: mitraayushi@gmail.com

Abstract:- Sentiment analysis or Opinion Mining or Emotion Artificial Intelligence is an on-going field which refers to the use of Natural Language Processing, analysis of text and is utilized to extract quantify and is used to study the emotional states from a given piece of information or text data set. It is an area that continues to be currently in progress in field of text mining. Sentiment analysis is utilized in many corporations for review of products, comments from social media and from a small amount of it is utilized to check whether or not the text is positive, negative or neutral. Throughout this research work we wish to adopt rule- based approaches which defines a set of rules and inputs like Classic Natural Language Processing techniques, stemming, tokenization, a region of speech tagging and parsing of machine learning for sentiment analysis which is going to be implemented by most advanced python language.

Keywords: Natural Language Processing; Sentiment Analysis; Opinion Mining; Stemming; Tokenization; Machine Learning

1. Introduction

Many new researchers nowadays have been influenced by on-line & public forum sites data analysis by applying slicing-dicing and mining algorithms to derive precious information out of raw data sources. Organizations these days assess their customers for their respective related products from social sites text dumps/raw logs [12]. Method of automatically classifying a user-generated text as positive text, negative text or neutral opinion is determined by Sentiment Analysis algorithms concerning an entity such as product, people, topic, event etc. Document level, Sentence level and Feature level are the three levels of classification in Sentiment analysis respectively [13]. This entire document is presented as a basic information unit to provide scope of classification addicted to positive or negative class at Document level specifically. Each sentence is classified initially into subjective or objective and then it is segregated as positive, negative or neutral category in Sentence level classification. In third level of Sentiment analysis classification that is Aspect or Feature level classification, it distinguishes by extracting product attribute details by analysing the source data [13]. Second approach in industry today is Lexicon based approach contains a dictionary of positive and negative words which is used to determine the sentiment polarity based on inclination of message from source dataset content that is source data set has more words in positive word repository or negative word repository. The combination of both Machine learning and lexicon based approach is then used by Hybrid based approach for classification.

1.1 Theoretical Background

Machine learning ways for sentiment analysis usually rely upon supervised classification methods, wherever tagged/labelled data is employed for the approach. In Fig-1 below we have tried to provide an overview of the entire architecture. Below, it depicts two methods (a) Training method, and (b) Prediction method. Within the Training method (a) model trains itself to adapt to a specific input (text) to the corresponding output (tag) which is based on sample data provided for training purpose, based on 80:20 principle. Here 80 % data is fed into the application with intention to train it. Rest 20 % is meant for the next phase that is prediction phase. Feature extractor function is to transfer text input from previous step into a feature vector, where the text-tag matric is built and then these feature vectors and tags (e.g. positive, negative, or neutral) are fed into the machine learning codes/algorithms that will generate a model.

ISSN: 2582-337X
Submitted: 15.08.2020
Accepted: 10.09.2020
Published: 18.09.2020

In the prediction phase (b) feature extractor work is to transform the unseen text inputs into feature vectors. These feature vectors are then fed into the model which will generate the predicted or expected tags (i.e. positive, negative, or neutral)that the model learnt for the 80% data sample in previous step.

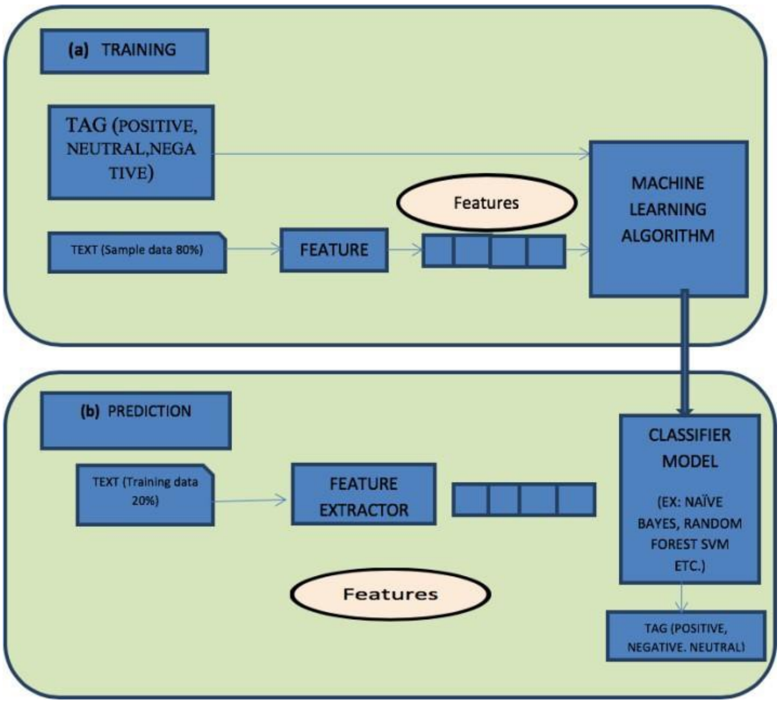

(a) TRAINING
TAG (POSITIVE, NEUTRAL,NEGA TIVE)
TEXT (Sample data 80%)
FEATURE
Features
MACHINE LEARNING ALGORITHM
(b) PREDICTION
TEXT (Training data 20%)
FEATURE EXTRACTOR
Features
CLASSIFIER MODEL
(EX: NAÏVE BAYES, RANDOM FOREST SVM ETC.)
TAG (POSITIVE, NEGATIVE. NEUTRAL)

Figure 1. Block diagram of Sentiment Analysis

2. Related Works

Rohit Joshi et al [4] used Natural Language Processing (NLP) Techniques to determine sentiments with the help of a tool Sentiment Analyzer that automatically extracts sentiments and is used to discover all references for the given subject with efficiency.

MinhoeHur et al [5] three machine learning based algorithms like artificial neural network, regression tree, and support vector regression were used to get non-linear relationship between the box-office collections based on Sentiments of movie review.

Agarwal et al [6] used numerous approaches and classifiers like lexicon based mostly approach, Naïve Bayes (NB) classifier algorithm Support Vector Machines (SVM) and MaximumEntropy(MaxEnt)

RafeequePandarachalil et al [7] a Twitter Sentiment analysis method was presented by using an unsupervised learning approach. SenticNet, SentiWordNet, and Sentis- langNet were the three Sentiment lexicons used to determine the polarity of tweets.

ChiragSangani [8] provides a collection of reviews to every topic that refers user opinions towards the topic and a many-to-many relation was established from reviewers to topics of interest.

Mudinas and Zhang [9] Hybrid techniques which were used are reliable techniques like lexicon based technique and performance as Machine learning based technique. Overall accuracy of the system was observed to be 82.3%

Koloumpis et al [10] Various features were used like unigrams, bigrams, n-grams, pos tagging and hash tags. The result which was found was of mixed classification.

Lei Zhang et al [11] They used the associated rule mining technique for extracting product features and differentiated between positive and negative reviews

ParvathyG, Bindhu JS et al [12] presented a hybrid approach consisting of machine learning techniques like artificial neural network, support vector machine, regression tree and rule based technique.

ISSN: 2582-337X
Submitted: 15.08.2020
Accepted: 10.09.2020
Published: 18.09.2020

3. Proposed Work

To perform Sentiment Analysis we have discussed approaches, and their accuracy results. We have shown the block diagram of Sentiment Analysis in Fig.1. Here we have presented test set, training set and prediction set and we have tried to provide an overview of the entire architecture, how machine learning model works on training the model first to gather experiences (usually 80% training data as input) and then prediction by the model using that gained experience in earlier step (usually 20% training data as input).It depicts two methods (a) Training method, and (b) Prediction method. Within the Training method (a) model learns to accompany a specific input (text) to the corresponding output (tag) which is based on the test samples for training. Feature extractor work is to transfer text input into a feature vector, then these feature vectors and tags (e.g. positive, negative, or neutral) are then fed into the machine learning- approaches that will generate a model. In the prediction method (b) feature extractor work is to transform the unseen text inputs into feature vectors. These feature vectors are then fed into the model which will generate the predicted or expected tags (i.e. positive, negative, or neutral) which is shown by CFG in the Figure 2.

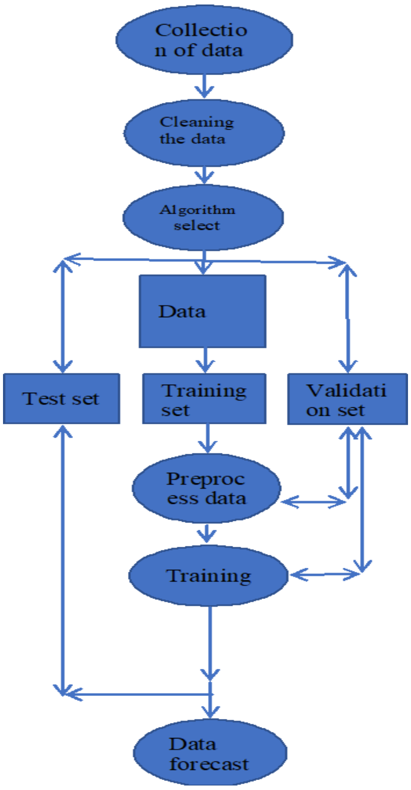


Figure 2. Workflow of the proposed framework.

4. Results and Discussion

To perform Sentiment Analysis we have discussed approaches, and their accuracy results in table 1. The main disadvantages of these approaches are:

a)To complete the work we will use Logistic Regression, ADA Boost, and Random Forest.

b)Most of the research articles have discussed about SVM(Support Vector Machines), Naïve Bayes, KNN(K-Nearest Neighbor), NLP(Natural Language Processing techniques).

Although in [1] the tweets were classified by many categories by introducing new features that powerfully identify the fine tune the polarity degree of a post and compared it with other Sentiment Analysis tools still need further research

Approach 1:

Classifier Algorithm & Accuracy details provided below:

a) Naïve Bayes [accuracy:0.7044117647058824]

b) SKlearnBernouliieNB [accuracy:0.7014705882352941]

ISSN: 2582-337X
Submitted: 15.08.2020
Accepted: 10.09.2020
Published: 18.09.2020

c) Sklearn SVC () [accuracy:0.7536764705882356]

Approach 2:

Classifier Algorithm & Accuracy details provided below:

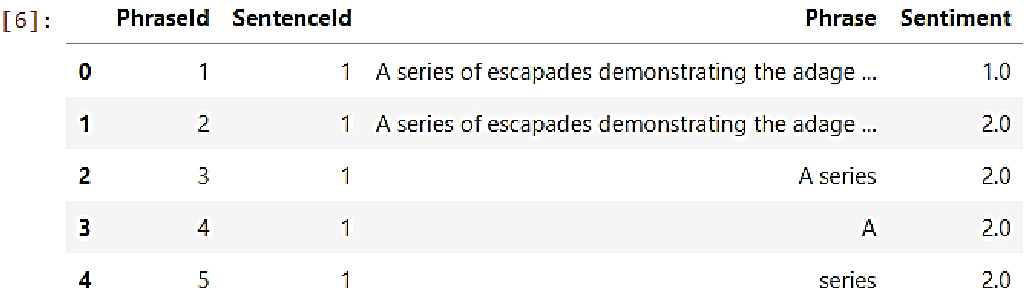
a) Decision Tree [accuracy:0.52]

b) Random Forest [accuracy:0.80]

c) KNN [accuracy:0.71]

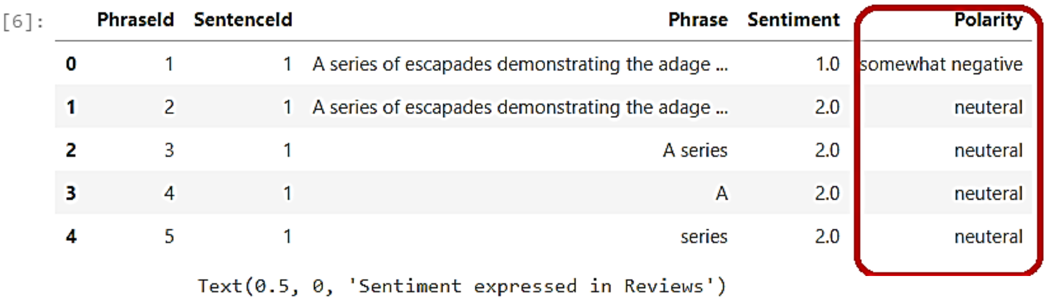
. Table 1. Accuracy Table

Feature	SVM	NBM
Unigrams	85.45	81.45
Unigrams	86.35	83.95
Bigrams	85.35	83.15
Adjectives	75.85	82.00



[6]:

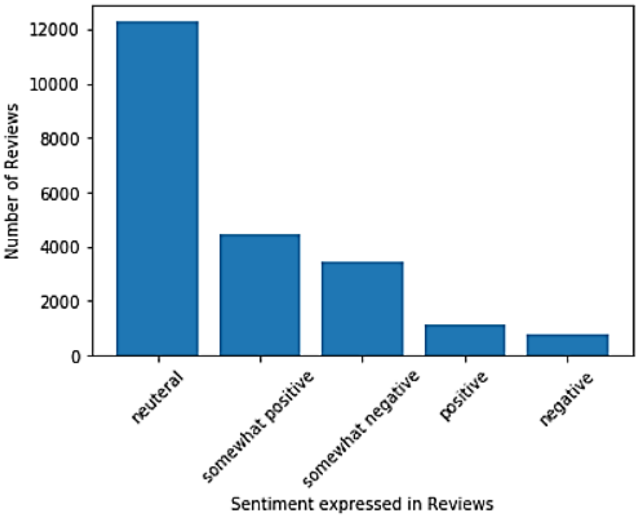
	PhraseId	SentenceId	Phrase	Sentiment
0	1	1	A series of escapades demonstrating the adage ...	1.0
1	2	1	A series of escapades demonstrating the adage ...	2.0
2	3	1	A series	2.0
3	4	1	A	2.0
4	5	1	series	2.0



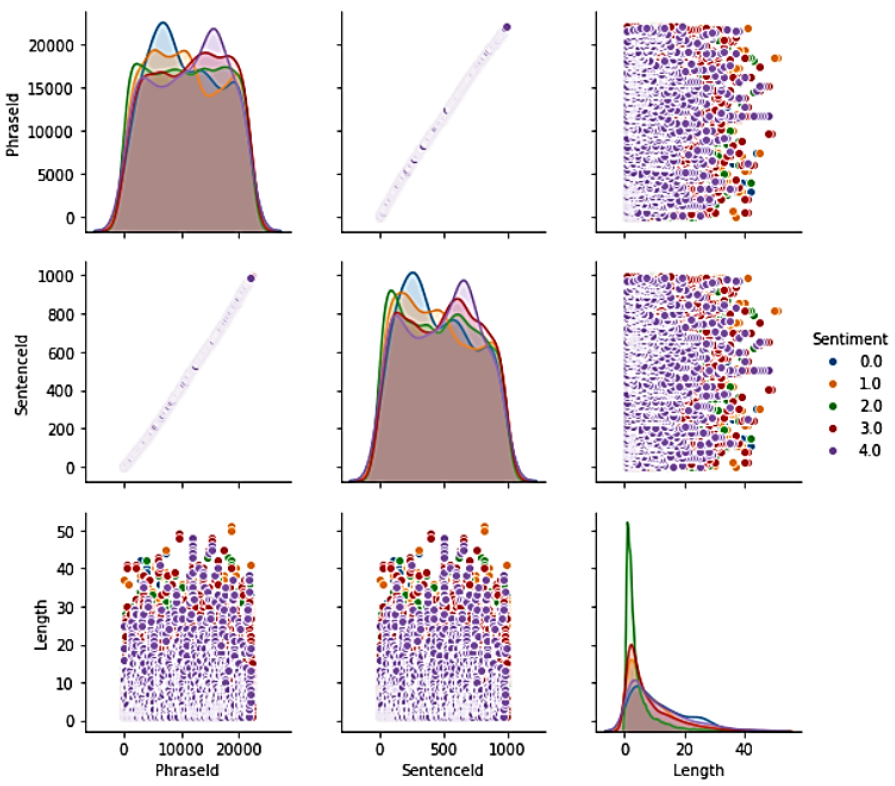
[6]:

	PhraseId	SentenceId	Phrase	Sentiment	Polarity
0	1	1	A series of escapades demonstrating the adage ...	1.0	somewhat negative
1	2	1	A series of escapades demonstrating the adage ...	2.0	neuteral
2	3	1	A series	2.0	neuteral
3	4	1	A	2.0	neuteral
4	5	1	series	2.0	neuteral

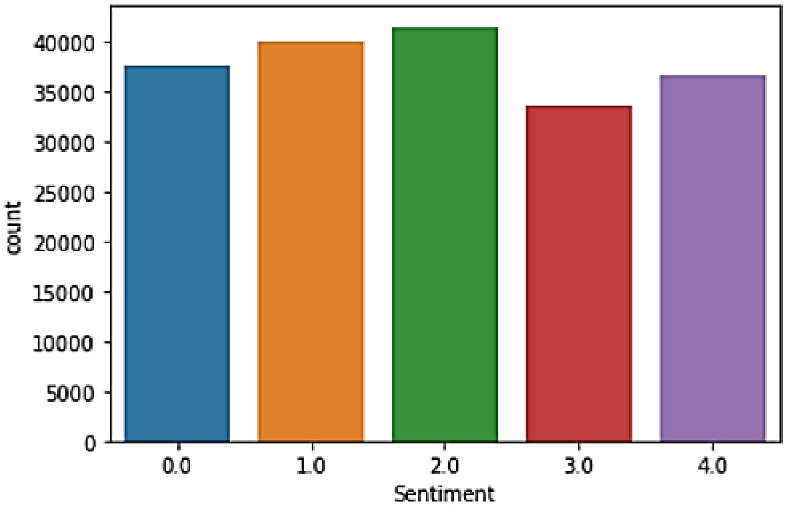
```
Text(0.5, 0, 'Sentiment expressed in Reviews')
```

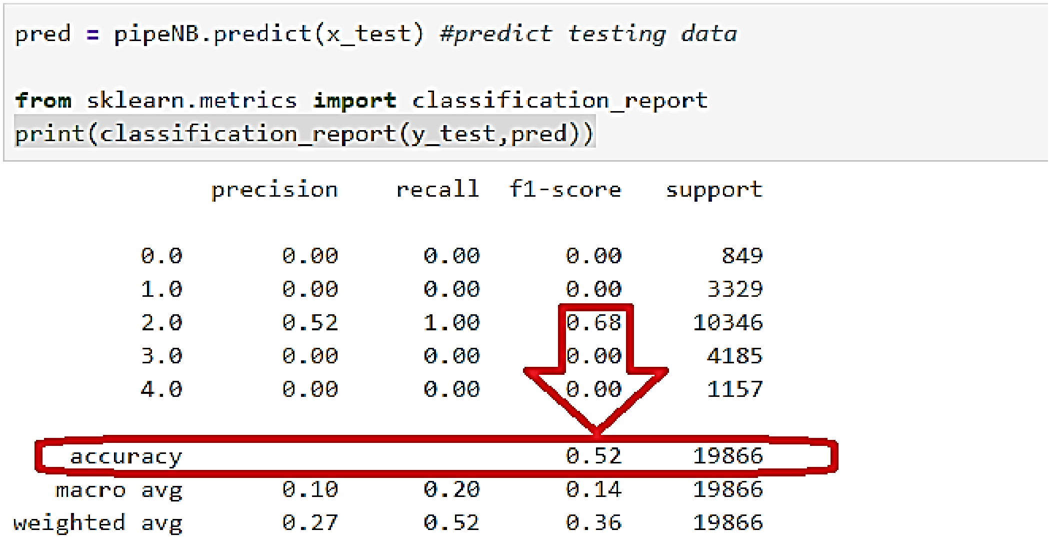


```
<seaborn.axisgrid.PairGrid at 0x7f5b69b56850>
```


PhraseId
SentenceId
Length
20000
15000
10000
5000
0
1000
800
600
400
200
0
50
40
30
20
10
0
0
10000
20000
0
500
1000
0
20
40
Sentiment
● 0.0
● 1.0
● 2.0
● 3.0
● 4.0

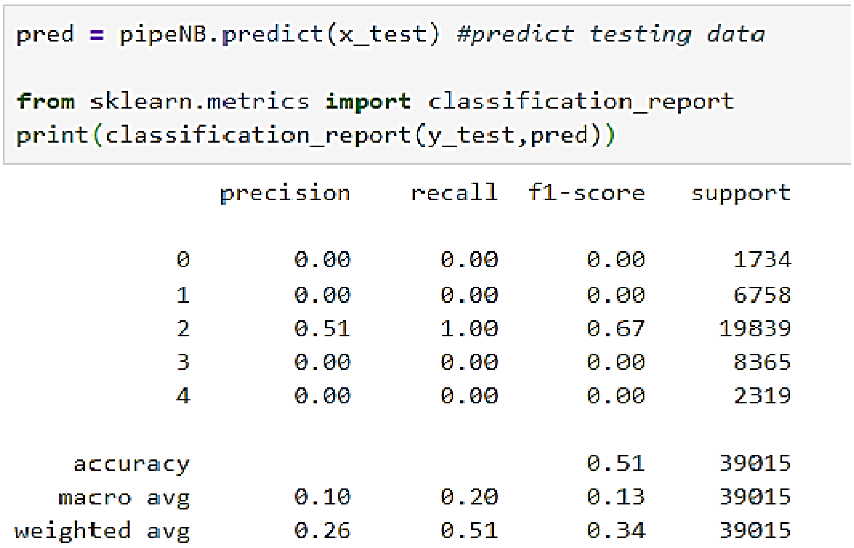
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8964e74c10>
```


count
40000
35000
30000
25000
20000
15000
10000
5000
0
0.0
1.0
2.0
3.0
4.0
Sentiment



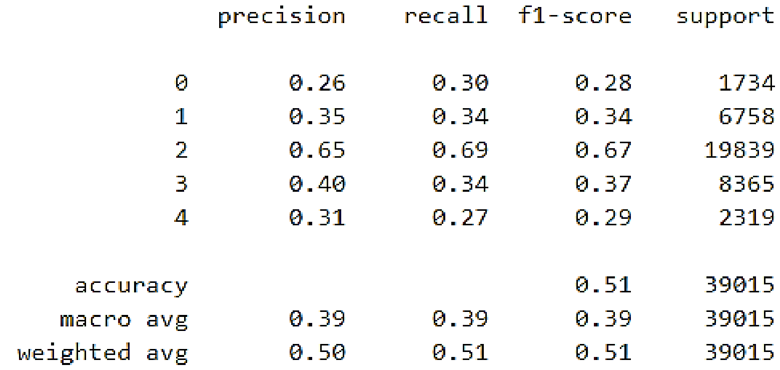
```
pred = pipeNB.predict(x_test) #predict testing data

from sklearn.metrics import classification_report
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

         0.0       0.00      0.00      0.00       849
         1.0       0.00      0.00      0.00      3329
         2.0       0.52      1.00      0.68     10346
         3.0       0.00      0.00      0.00      4185
         4.0       0.00      0.00      0.00      1157

    accuracy                           0.52     19866
   macro avg       0.10      0.20      0.14     19866
weighted avg       0.27      0.52      0.36     19866
```

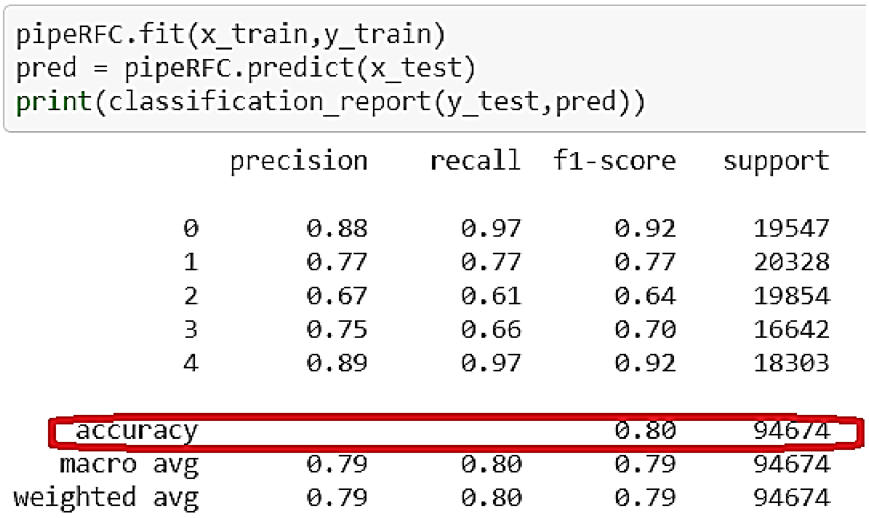


```
pred = pipeNB.predict(x_test) #predict testing data

from sklearn.metrics import classification_report
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.00      0.00      0.00      1734
           1       0.00      0.00      0.00      6758
           2       0.51      1.00      0.67     19839
           3       0.00      0.00      0.00      8365
           4       0.00      0.00      0.00      2319

    accuracy                           0.51     39015
   macro avg       0.10      0.20      0.13     39015
weighted avg       0.26      0.51      0.34     39015
```

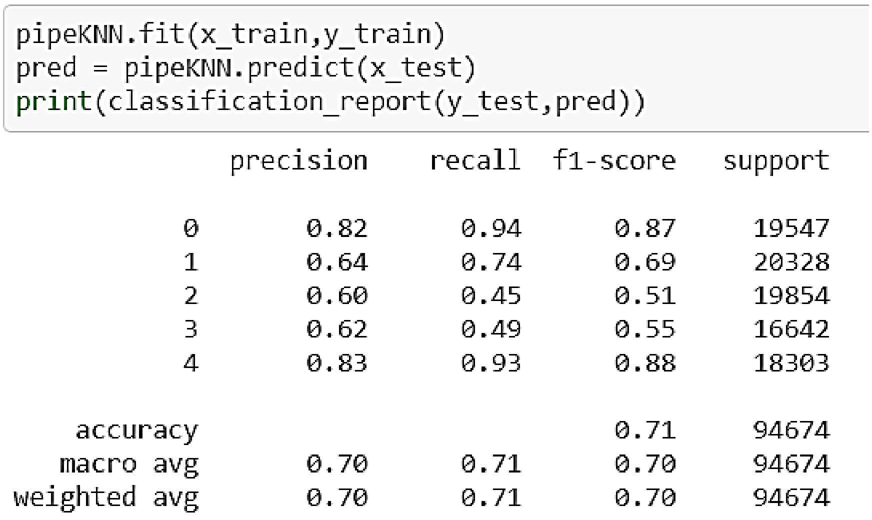


```
              precision    recall  f1-score   support

           0       0.26      0.30      0.28      1734
           1       0.35      0.34      0.34      6758
           2       0.65      0.69      0.67     19839
           3       0.40      0.34      0.37      8365
           4       0.31      0.27      0.29      2319

    accuracy                           0.51     39015
   macro avg       0.39      0.39      0.39     39015
weighted avg       0.50      0.51      0.51     39015
```



```
pipeRFC.fit(x_train,y_train)
pred = pipeRFC.predict(x_test)
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.88      0.97      0.92     19547
           1       0.77      0.77      0.77     20328
           2       0.67      0.61      0.64     19854
           3       0.75      0.66      0.70     16642
           4       0.89      0.97      0.92     18303

    accuracy                           0.80     94674
   macro avg       0.79      0.80      0.79     94674
weighted avg       0.79      0.80      0.79     94674
```



```
pipeKNN.fit(x_train,y_train)
pred = pipeKNN.predict(x_test)
print(classification_report(y_test,pred))
```

```
              precision    recall  f1-score   support

           0       0.82      0.94      0.87     19547
           1       0.64      0.74      0.69     20328
           2       0.60      0.45      0.51     19854
           3       0.62      0.49      0.55     16642
           4       0.83      0.93      0.88     18303

    accuracy                           0.71     94674
   macro avg       0.70      0.71      0.70     94674
weighted avg       0.70      0.71      0.70     94674
```

5. Conclusion

By using lexicon based approach, machine learning based approach or hybrid approach Sentiment analysis will be performed. In this related field already researches have been made to find the accurate accuracy still their results seems to be inefficient. The strength of the sentiment classification depends on the scale of the lexicon (dictionary) because the size of the lexicon will increase this approach and becomes more incorrect and time consuming. We will be using NLTK (Natural Language Toolkit) feature in python for further implementation sample movie review data. This will focus upon using in-built classifier models from NLTK package in python and compare their accuracy for a given dataset.

References

[1] El Alaoui, Imane, Youssef Gahi, RochdiMessoussi, YounessChaabi, Alexis Todoskoff, and AbdessamadKobi. "A novel adaptable approach for sentiment analysis on big social data." Journal of Big Data 5, no. 1 (2018).

[2] Jaspreet Singh, Gurvinder Singh and Rajinder Singh, Hum, Cent, Comput, Inf ,Sci,(2017) DOI 10.1186/s 13673-017-0116-3.

[3] OnamBharti, Mrs, and Monika Malhotra."SENTIMENT ANALYSIS." (2016).

[4] Joshi, Rohit, and RajkumarTekchandani. "Comparative analysis of Twitter data using supervised classifiers." In 2016 International Conference on Inventive Computation Technologies (ICICT), vol. 3, pp. 1-6.IEEE, 2016.

[5] Hur, Minhoe, Pilsung Kang, and Sungzoon Cho. "Box-office forecasting based on sentiments of movie reviews and Independent subspace method." Information Sciences 372 (2016): 608-624.

[6] Agarwal, Basant, SoujanyaPoria, Namita Mittal, Alexander Gelbukh, and Amir Hussain. "Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach." Cognitive Computation 7, no. 4 (2015): 487-499.

[7] Pandarachalil, Rafeeque, SelvarajuSendhilkumar, and G. S. Mahalakshmi. "Twitter sentiment analysis for large-scale data: an unsupervised approach." Cognitive computation 7, no. 2 (2015): 254-262.

[8] Sangani, Chirag, and SundaramAnanthanarayanan. "Sentiment analysis of app store reviews." Methodology 4, no. 1 (2013): 153-162.

[9] Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." In Proceedings of the first international workshop on issues of sentiment discovery and opinion mining, p. 5.ACM, 2012.

[10] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." In Fifth International AAAI conference on weblogs and social media. 2011.

[11] Lei Zhang, Bing Liu, Suk Hwan Lim, Eamonn O' Brien-Strain-Proceedings of the 23rd international conference computational linguistics, 2010.

[12] Parvathy G, Bindhu JS (2016) A probabilistic generative model for mining cybercriminal network from online social media: a review. Int J ComputAppl 134(14):1-4.doi:10.5120/ijca2016908121 Google Scholar.

[13] Vohra, S. M., and J. B. Teraiya. "A comparative study of sentiment analysis techniques." Journal JIKRCE 2, no. 2 (2013): 313-317.

Authors Biography

Ayushi Mitra
College of Engineering and Technology, Bhubaneswar,
Bhubaneswar, Odisha, India.
Email: mitraayushi@gmail.com

ISSN: 2582-337X
Submitted: 15.08.2020
Accepted: 10.09.2020
Published: 18.09.2020