Sentiment Analysis on Reviews of Scientific Papers Using Machine Learning Approach

Nirdesh Yadav†
College of Computer Science and Technology
Lovely Professional University
Phagwara India
My43557@gmail.com

Rajkumar Murao
College of Computer Science and Technology
Lovely Professional University
Phagwara India
Rajkumar.11812510@gmail.com

Abstract: Sentiment analysis is that the method of distinguishing and managing opinions expressed during a piece of text to examine whether or not the author's perspective towards explicit topics or the merchandise is positive, negative, or neutral. This paper offers access to the machine learning algorithms and Scikit-learn in sentiment analysis of the Review dataset. To perform this, we've got analyzed Review datasets created publically accessible by NLTK Corpora Associate in Nursingd created an economical feature by employing a feature extraction technique for the extraction of a dataset. we have a tendency to train and take a look at datasets victimisation numerous machine learning classifiers like logistical Regression, Sklearn, IfidVectorizer, Pipeline, and Confusion Matrix. The experimental results demonstrate that sklearn, and LogisticRegression classifier reached accuracy as high as 75%.

Introduction: **Nowadays** various researchers are influenced by online appliances and websites for data analysis by applying slicing-dicing and mining algorithms in order to derive precious information from raw data sources. Organizations are accessing their customer's views or emotion for their respective related products from social sites. The way of classifying a user-generated text as positive text, negative text, or neutral opinion can determine by using Sentiment Analysis algorithms concerning an entity such as product, people, topic, event, etc. Sentiment analysis is a process of analysis of text utilized to extract quantify which is used to study the emotional states from a given dataset. It also includes various other aspects of this project such as vectorization, logistic regression, Sklearn model, etc. Sklearn offers a pool of efficient tools for machine learning and statistical modelling including, clustering, classification, dimensionality reduction and regression via a consistence interface in Python. Sklearn models tackle the selection of efficient tools for machine learning and statistical modeling. Logistic Regression is a predictive analysis algorithm used for the classification problems. It is based on the concept of probability. Logistic Regression uses a more complex cost function called 'Sigmoid function'. Logistic regression is a supervised learning techniques. Vectorization is basically a technique to implement arrays in machine learning models without the use of loops which can minimize the running time and execution time of code efficiently. This analysis is mainly focused on finding the accuracy of the models by analyzing training and testing data for calculation on precision, recall, and accuracy. It also shows the result

of sentiment analysis for the Review datasets.

2. Methodology

This section describes the steps of sentiment analysis on research papers reviews using machine learning techniques and Scikit-learn. To begin with, we explain the source of the dataset. In addition, we perform pre-processing work to remove all the unwanted things in the dataset. Moreover, we use the Sklearn model techniques to extract the important features from the clean dataset, and finally, we discuss all the machine learning techniques and work done on the dataset in order to get better accuracy. As we all know, there are different procedures and method to work with the data, in this case reviews, fetch the reviews, reviews pre-processing, classify reviews which means obtain the polarity of the data, and finally return the results.

Algorithm :Extracts data
Begin
Input Query String
Until the data is retrieved from website, Do
Filter paper review data

Remove duplicate data
For each review, Do:
Procedure Pre-processing(review):
Eliminate all the private symbols(#topic, @user name, retweet (RT))
Remove URL
Remove all numbers, symbols, punctuations, and Emoticons
Removal of Stop Words
Stemming
Tokenization
Return reviews

End Procedure
Procedure Feature Extraction (cleaned reviews):
Extract Features in a format suitable to machine learning algorithms
End Procedure
Procedure Sentiment Classification (Feature):
Extract Features in a format suitable to machine learning algorithms
End Procedure
Procedure Sentiment Classification (Feature):
Classify reviews using machine learning techniques

2.1 Creating Dataset

The dataset is taken from the paper reviews websites. The dataset has column with the name text which consist the reviews of a papers and the other column with the name of preliminary decision with consist the possibility of whether paper going to accept , reject or probably reject. The total number in dataset is 407 reviews consisting of 260 positive (accept) reviews and 118 negative (reject) reviews and the remaining with probably reject. Statistics of dataset is shown in table 1

Total	**Accept**	**Reject**	**Probably reject**	**No decision**
407	260	118	20	2

Table 1 Format of dataset

2.2 Pre-processing

Pre-processing is a major step for handling raw data. A review is a short message full of noise such as irrelevant emoticons, symbols, stop words, misspellings, and slang words. Raw data scraped from sites generally result in a noisy dataset. These types of noisy characteristics affect the performance of the machine learning modules. Raw data has to be cleaned in order to create a dataset that can learn from machine learning algorithm. So to avoid this, some pre-processing is applied before extracting features of dataset and using the specific machine learning algorithms. Pre-processing of data is done in order to standardize the dataset and reduce its size. The noisy data or unwanted characters are removed by lemmatization, stop words removal, stemming, etc., Some of the pre-processing has perform to convert the capital word from reviews to lower case, remove spaces and quotes (“Machine learning “) Replacing two or more dots (.)

With space, from the ends of reviews, Replacing 2 or more spaces with single space, etc., For example, consider the sentence, “This is very important line of the CODE!” After pre-processing the resultant sentence is “very important line code”

2.3 Feature Extraction

The feature extraction module used to extract the traits in suitable way to be used directly in machine learning algorithms from datasets containing raw data of formats such as a sequence of symbols, text and image. Therefore we use Term Frequency-Inverse Document Frequency (TF-IDF). This is the techniques used to quantify the words in a set of document. We calculate the weightage of each word to specify its importance in the document and corpus. TF-IDF is a statistical measure used to calculate how important a word is to a document in a dataset. Each word is given a weight in the document [4], and different meanings are extracted from the processed dataset, such as verbs, nouns, and adjectives. Later, these different aspect are used to evaluate the positive or negative polarity in a sentence, which is helpful for determining the opinion of the individuals using models such as unigrams, bigrams, or n-grams. We used a TF-IDF Vectorizer Python module of Scikit-learn [5] to extract TF-IDF.

2.4 Machine Learning Techniques

There are several types of classifiers used for machine learning techniques to identify sentiment analysis of social media or websites data. Machine learning offers numerous solution to the sentiment classification problem. In order to create our classifier, we used a Python library called Scikit-learn. It is a powerful and very useful open-source machine learning package in Python that provides many classification algorithms.

The classification models selected for this study are:

2.4.1 Scikit-learn Library: Scikit-learn (Sklearn) is the most useful and precise library for machine learning in Python. It was originally known as scikits.learn. It was initially developed by David Cournapeau. It offers a pool of efficient tools for machine learning and statistical modelling including, clustering, classification, dimensionality reduction and regression via a consistence interface in Python. David Cournapeau developed skicit-learn as a Google summer of code project in 2007. This library, which is mostly written in Python, is built upon Matplotlib, NumPy, and SciPy. Later, in 2010, four scientist from French Institute for Research in Computer Science and Automation, Alexandre Gramfort, Fabian Pedregosa, Gael Varoquaux, and Vincent Michel, took this project even further and on 1st February 2010, they made the first public release (v0.1 beta).

2.4.2 Sklearn Train-Test Split: train test split is a model validation procedure that allows you to simulate how a model would perform on new/unseen data. Here is how the procedure works.

Make sure your data is arranged into a format acceptable for train test split. In scikit-learn, this consists of separating your full dataset into Features and Target.

Split the data in set of training and testing. This consists of randomly selecting about

75% (you can vary this) of the rows and putting them into your training set and putting the remaining 25% to your test set. Now data will split into two part, later make the four variable for better training and testing with the names of ("X_train", "X_test", "y_train", "y_test") for a particular train test split.
Train the model on the training set. This is "X_train" and "y_train".
Test the model on the testing set ("X_test" and "y_test") and evaluate the performance.
It is an easy and fast procedure to perform, the result allow us to compare the performance of machine learning algorithms for our predictive modeling problem. Although simple to use and interpret, there are conditions when this method should not be used, such as in a case of a small dataset and situations where additional configuration is needed, such as when it is used for classification and the dataset is not balanced.

2.4.3 Logistic Regression Algorithm: Logistic Regression could be a prognostic analysis algorithmic program used for the classification issues. it's supported the construct of chance. Logistic Regression uses a additional complicated price perform referred to as 'Sigmoid function'. The hypothesis of provision regression tends it to limit the value perform between zero and one. LogisticRegression() could be a perform wont to predict the accuracy of the sentiment classification from the input of term frequency whole number vectors

$$0 \leq h\emptyset z \leq 1$$

$$Sigmoid\ function\ \sigma(z) = \frac{1}{1+e^{-z}}$$

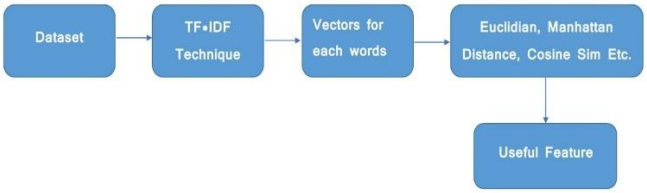
Logistic Regression is basically used to identify the reviews as positive(accept) and negative(reject) from the inputs of TFV (term frequency vectors) and then calculate the accuracy of project with the parameters such as penalty: default: l2, C: 1.0, Class weight= None, Tol= 0.0001, Maximum iteration= 1000.

2.4.4 Support vector machine Algorithm: Support vector machine can be used to classify the tweets as positive and negative from the inputs vectors and then calculate the accuracy of sentiments with the help of term frequency. LinearSVC() is used to predict the accuracy of the classification data based on the marginal probability of the input vectors. The parameter values set are penalty: default: l2, C: 1.0, Class_weight= none, Tol= 0.0001, Max_iter= 1000.

2.4.5 TF-IDF: Term Frequency (TF) is used to calculate the number of occurrences of a word in a document. TF indicates the significance of a particular term within the overall dataset.Term frequency is calculated for each word in the tweet and this frequency vector is used to train the model.

$$TF = \frac{number\ of\ the\ term\ x\ appear\ in\ the\ document}{total\ number\ of\ term\ in\ document}$$

The inverse document frequency calculate how much information the word provides. It measures the weight of a given word in the entire document. IDF shows the occurrence of ant word in the document, like how rare or frequently a word appear in the document. The pipeline of TF-IDF in shown in image[1]



2.4.6 Confusion Matrix: Scikit learn confusion matrix is known as a technique to calculate the performance of classification. The confusion matrix is also used to summarise or predict the result of the classification problem. You can plot confusion matrix using the confusion_matrix () method from sklearn.metrics package. Confusion matrix provides the visualization of the performance of the classification machine

learning models. With the help of this visualization, we can get a better idea of how our ML model is performing.

3.Results and Discussion

This section presents the experimental result of the classification accuracy of machine learning algorithms for sentiment analysis. The Sklearn algorithms were tested on scientific paper reviews dataset. Term frequency is calculated for each word in the reviews and this frequency is used to train the model. Pre-processing and feature extraction techniques were done on this labelled dataset. An experiment was performed on scikit-learn and Logistic Regression machine learning classifiers. To begin with, we started with checking the missing values in the dataset, that is where isnull () function comes handy. This operation is important, in order to maintain high accuracy of the model. The missing values interfere with the models feature extraction. The result of this operation shows in image[2]

Out[277]:

	Count	Percentage
text	0	0.0
id.1	0	0.0
confidence	0	0.0
preliminary_decision	0	0.0
id	0	0.0

Image[2]

Image[2] shows the percentage of missing values of each column in the dataset.

Later, we used matplotlib library to checking the distribution of the polarity of the dataset. The dataset holds four types of decisions: accept, reject, probably reject, and no decision.

```
Percentage for default

accept             64.7059
reject             29.9020
probably reject     4.9020
no decision         0.4902
Name: preliminary_decision, dtype: float64
```

Image[3]

The image[3] shows percentage of availability of each decision in the dataset .

	id	preliminary_decision	confidence	id.1	text	cleaned_review	cleaned_review_new
0	1.0	accept	4.0	1.0	- The article deals with a contingent and very...	the article deals with a contingent and very ...	the article deals with a contingent and very ...
1	nan	accept	4.0	2.0	The article presents practical recommendations...	the article presents practical recommendations...	the article presents practical recommendations...
2	nan	accept	5.0	3.0	The topic is very interesting and a guide to i...	the topic is very interesting and a guide to i...	the topic is very interesting and a guide to i...
3	2.0	accept	4.0	1.0	An experience of using ICT for academic collab...	an experience of using ict for academic collab...	an experience of using ict for academic collab...
4	nan	accept	4.0	2.0	nan	nan	nan
5	nan	accept	4.0	3.0	The authors describe a methodology for collabo...	the authors describe a methodology for collabo...	the authors describe a methodology for collabo...
6	3.0	accept	4.0	1.0	This work proposes a new approach based on [25...	this work proposes a new approach based on to...	this work proposes a new approach based on to...
7	nan	accept	3.0	2.0	This paper aims to show new deployment alterna...	this paper aims to show new deployment alterna...	this paper aims to show new deployment alterna...
8	nan	accept	3.0	3.0	The paper is well structured. It follows a log...	the paper is well structured it follows a logi...	the paper is well structured it follows a logi...

Image[4]

Image[4] is the result after performing the pre-processing, removing all the unwanted data available in the dataset like Punctuation, capital letters, unwanted column, spaces, dots, pronouns, etc.

Moving further, we used scikit's test_train_split method to split data into two part for training and testing. We provides 0.2 of data into testing and 0.8 into training. Later, we used logistic regression and tf-idf in a pipeline and created a confusion matrix and runs the prediction on training data. In the last, we import accuracy_score, precision_score, and recall_score for calculating their values. The result of this process in shown in image[6].

```
Accuracy :  0.7439024390243902
Precision :  0.9684035476718403
Recall :  0.7439024390243902
```

Image[6]

4. CONCLUSION AND FUTURE ENHANCEMENT

Sentiment analysis is a process to identify the opinion of a text. People post comments in social media mentioning their experience about an event and are also interested to know if the majority of other

people had a positive or negative experience on the same event. The goal is to calculate the sentiment accuracy of sentences that were extracted from the text of reviews of scientific papers. The sentiment analysis of the reviews helps to find whether the sentiment of the reviews on particular products, events, etc., is positive or negative. The most challenging task in sentiment analysis is to identify an opinion word as either positive or negative. By changing the parameter of the machine learning algorithms is possible to get the best accuracy. This work calculate the performance of machine learning approaches includes Scikit-learns and Logistic regression. The Sklearn model with the help of Logistic Regression has achieved an accuracy of approximately 75%. Logistic regression performs better when compared to other supervised machine learning algorithms for reviews sentiment analysis. The future work can be of analysing the fluctuation in the performance of sentiment analysis algorithm when multiple features are considered. A further active learning techniques like expected error reduction, pool-based sampling, uncertainty sampling, and so on shall be utilized to detect sentiments and to increase the confidence of decision makers.

5. REFERENCES

[1] Poornima. A, K. Sathiya Priya ," A Comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques" on 6th International Conference on Advanced Computing & Communication Systems (ICACCS) ,2020

[2] Shihab Elbagir†, Jing Yang, "Sentiment Analysis of Twitter Data Using Machine Learning Techniques and Scikit-learn" on ACAI'18, December, 2018, Sanya, China.

[3] K. Tao, C. Hauff, G.-J. Houben, F. Abel, and G. Wachsmuth, "Facilitating Twitter data analytics: Platform, language and functionality," in Big Data (Big Data), 2014 IEEE International Conference on, 2014, pp. 421–430

[4] C.D. Manning, P.Raghavan and H. Schutze , "Introduction to Information Retrieval",Cambridge University Press,PP.234-265,2008.

[5] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," Math. Program., vol. 162, no. 1–2, pp. 83–112, 2017.

[6] Abdullah Alsaeedi, Mohammad Zubair Khan, "A Study on Sentiment Analysis Techniques of Twitter Data", *International Journal of Advanced Computer Science and Applications*, Vol.10, No.2, 2019.

[7] Suchita V Wawre, Sachin N Deshmukh, "Sentiment Classification using Machine Learning Techniques", *International Journal of Science and Research (IJSR)*, Vol.6, 2015.

[8] Ali Hasan, Sana Moin, Ahamad Karim and Shahaboddin Shamshirband, "Machine Learning-Based Sentiment Analysis for Twitter Accounts", *Journal mca*, 16 January 2018, Accepted 24 February 2018, Published: 27 Febryary 2018.

[9] Vishal A. Kharde, S. S. Sonawane, "Sentiment Analysis of Twitter Data: A survey of Techniques", *International Journal of Computer Applications (0975-8887)* Volume 139, No.11, April 2016.

[10] Suchita V Wawre, Sachin N Deshmukh, "Sentimental Analysis of Movie Review using Machine Learning Algorithm with Tuned Hyperparameter", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.4, Issue 6, June 2016.

[11] Zohreh Madhoushi, Abdul Razak Hamdan, Suhaila Zainydin, "Sentiment Analysis Techniques in Recent Works", *Science and Information Conference* 2015 July page no. 28-30, 2015.

[12] Bac LeHuy, Nguyen, **"**Twitter Sentiment Analysis Using Machine Learning Techniques", *conference paper, Advanced computaional methods for*

knowledge engineering,volume 358, pp 279-289, 2015.
[13] Alec Go, Richa Bhayani, Lei Huang, "Twitter Sentiment Classi□cation using Distant Supervision", *Semantic Scholar*, page no. 57-61, published 2009.
[14] Richa Sharma, Shweta Nigam, Rekha Jain, "Polarity Detection at Sentence Level", *International Journal of Computer Applications (0975 – 8887)*, Volume 86 – No 11, January 2014,29.