TOWARDS DATA SCIENCE - SENTIMENT ANALYSIS ON BIOMEDICAL JOURNAL

[1]P.Sudhasini, [2]Dr.B.Ashadevi

[1]Research Scholar ,[2]Assistant Professor
[1]Department of Computer Science,
[1]Mother Teresa Women's University, Kodaikanal,Tamilnadu, India
[2]Department of Computer Science,
[2]M.V. Muthaiah Govt. Arts College for women, Dindugul, Tamilnadu, India

Abstract : In the modern era, technology has been placed a major role on communication through internet on various fields. On part of education field to promote the existing growth of studies, online education has given opportunity to explore much knowledge. At the same time consistent level of education standard is depends on the research activities. Now days, the research focus are seemingly less, most of them publishing the papers only for the academic need and purpose of updating in their curriculum vitae. In this research, specifically get in to deep keen on journal article semantic analysis of citation for the necessity to identify the negative or positive polarity of citation comment for the existing research work. It gives impact to the researchers whether the existing work have been commented in positive way to carry down the work for future or negatively cited on part of contradiction or dissimilarity or requirements of quantifying things in clarity according to the proposed and existing work. It is not just for the author to get the count of articles has been cited his own article to feed in their resume but have to make sure the sense of comment suggested to his work. This sentiment analysis on research articles should be the essential part on paving the way to reason out the role and purpose of existing research work, that can be the healthiest mode to the author on knowing the recent research activities as well as to improve the existing work that has to be matched with new age of scientific research. In this research work, proposed sentiment analysis especially for clinical and medical journal of articles to find the negative citations. The proposed method implemented using Afinn Lexicon, Sentiword Lexicon to check the negative polarity of citation comments and finally used NaiveBayseCalssifier algorithm to classify the new arrival of comments with existing trained dataset.

***IndexTerms* – Data science, Natural language processing, Citation, Sentiment Analysis**

I. INTRODUCTION

In technology world communication through internet is rapidly growing and spreading to different fields to share their opinion. The sentiment analysis is playing a vital role on analyzing users perspective in indirect mode of interaction in the internet like sharing information of emotions, behavior patterns, positive and negative quotes. According to this research work, the opinion of author can be expressed by the way of research articles either to promote the present study or to give suggestions to the existing work to determine the authors attitude towards a particular or relevant research work to analyse the negative or positive sentiment on the comment made by the authors. On empowering education to higher standards in a consistent way the level of research promotion should be modest and ethical for the growth of healthiest research activities. The Sentiment analysis of negative citation can promote the way of qualitative research movement rather than quantitative research focus only on publications and credits. Qualitative research output lined up an vital role in ranking institutions and countries, allocating research grants, making hiring decisions and planning scientific priorities

II. RELATED WORK

The author Awais Athar (2011) focus on the problem of automatic identification of positive and negative sentiment polarity in citations to scientific papers. Using a newly constructed annotated citation sentiment corpus and explore the effectiveness of existing and novel features, including n-grams, specialized science-specific lexical features, dependency relations, sentence splitting and negation features. The results show that 3-grams and dependencies perform best in this task; they outperform the sentence splitting, science lexicon and negation based features. Christian Catalini et.al(2015) found that negative citations concerned higher-quality papers, were focused on a study's findings rather than theories or methods, and originated from scholars who were closer to the authors of the focal paper in terms of discipline and social distance, but not geographically. Kumar ravi et.al(2018) , proposed ensemble feature engineering method for deep learning that uses embedding of text and dependency relationships because of data scarcity of negative citations on article.

S. Anupkant , On sentence based collection of dataset author they implemented Logistic regression on training set resulted in 0.98 precision obtained on a classification model; hence the obtained regression model was applied on test set data to reveal positive and negative opinions. Diana C. Cavalcanti have used sentiword lexicons score to rate the degree of positivity and negativity for each adjective Relevance scores were then computed to rank citations according to the sentiment expressed in the text corresponding to each citation. The ranking generated by sentiment scores had an improved accuracy. G.Parthasarathy et.al, proposed citation architecture by analyzing the polarity of adjectives on comments. Also implemented various classifiers to derive the accuracy of polarity value on judging the negative comments. XIAOMEI BAI et.al calculated the impact of scholarly papers by PageRank and HITS algorithms, based on a credit allocation algorithm which is utilized to assess the impact of institutions fairly and objectively.

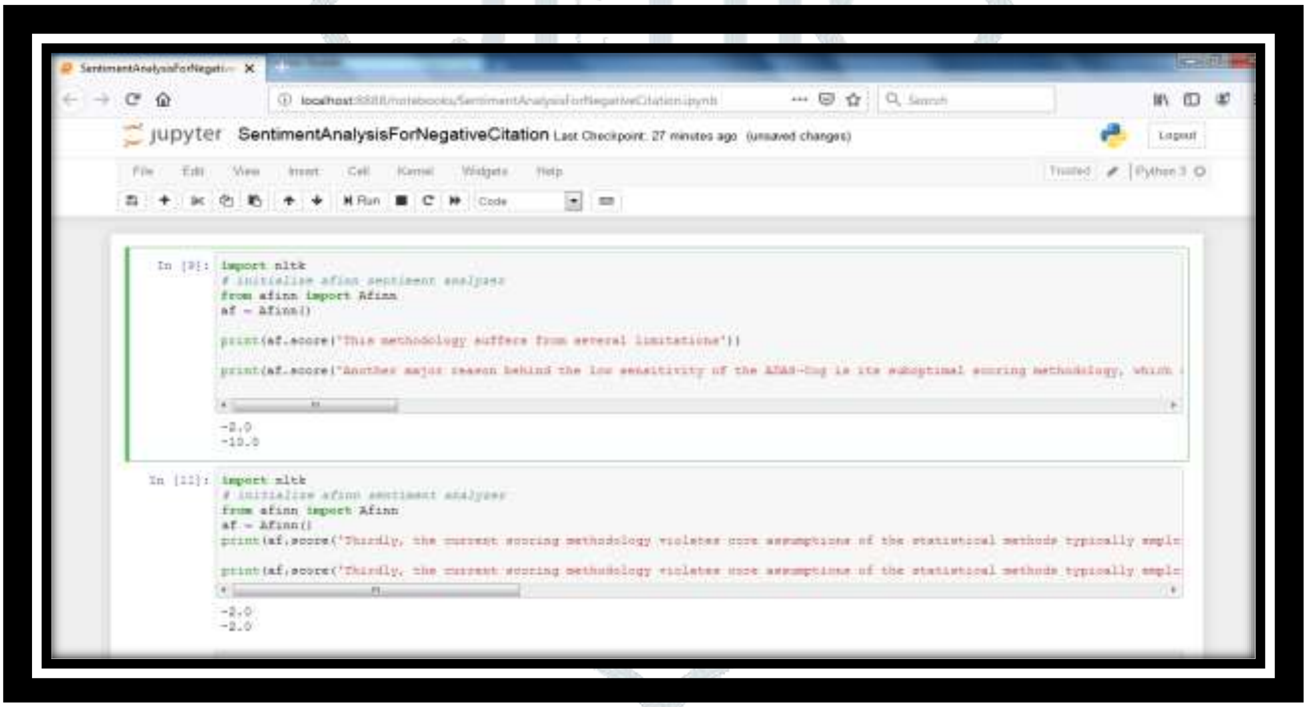
II. DATA COLLECTIONS

The data has been collected from PMC PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM) and BMJ (British Medical Journal) is a weekly peer-reviewed medical journal. It is one of the world's oldest general medical journals , High impact medical research journal. Champion of better research, clinical practice & healthcare policy since 1840. Nearly 50 records of negative citation of journal articles has been identified by reading the documents manually also implemented automated retrieval of information imbibed with sentiment analysis algorithm.

III. PROPOSED METHODOLOGY:

3.1. Using Afinn & Sentiword Lexicon to Analyse Negative Polarity of Journal Papers

The AFINN lexicon is one of the simplest and most popular lexicons that can be used extensively for sentiment analysis. The current version of the lexicon is *AFINN-en-165. txt* and it contains over 3,300+ words with a polarity score associated with each word. At the Initial stage to check the negativity of sentences have used Afinn lexicon with the version of AFINN-en-165. The sentiment of negativity score has been viewed as <=-1 (negative) and >=1(positive). According to statements consists of words relevant to the strength of negativity and positive polarity the score will be vary. Below is the example (Figure 1) screen shot of working with Afinn Lexicon.



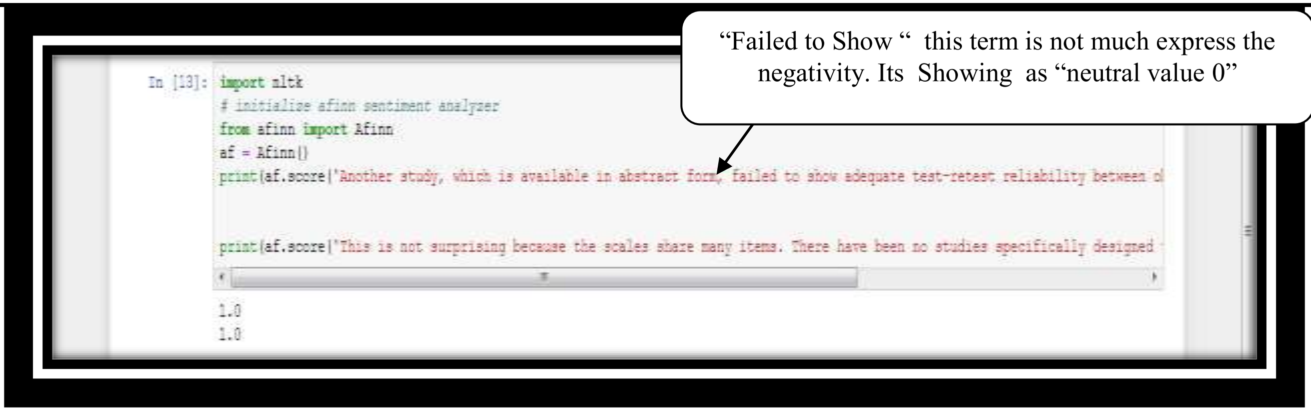
SentimentAnalysisForNegati
localhost:8888/notebooks/SentimentAnalysisForNegativeCitation.ipynb
Search
jupyter SentimentAnalysisForNegativeCitation Last Checkpoint: 27 minutes ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code

```
In [9]: import nltk
        # initialize afinn sentiment analyzer
        from afinn import Afinn
        af = Afinn()

        print(af.score('This methodology suffers from several limitations'))

        print(af.score("Another major reason behind the low sensitivity of the ADAS-Cog is its suboptimal scoring methodology, which

        -2.0
        -10.0

In [11]: import nltk
         # initialize afinn sentiment analyzer
         from afinn import Afinn
         af = Afinn()
         print(af.score('Thirdly, the current scoring methodology violates core assumptions of the statistical methods typically empl

         print(af.score('Thirdly, the current scoring methodology violates core assumptions of the statistical methods typically empl

         -2.0
         -2.0
```

Figure 1: Score of Negative polarity Statement using Afinn Lexiocon

Though its showing the negative polarity based on strength of words have been used in the citation, at some part of semantic clarity should be noted like not all the citations can give the explicit view of comments to show the negativity. See Figure 2 So, Based on the below result, decided to train the data to classify the citation as negative or polarity using NaiveBayseClassifier algorithm.

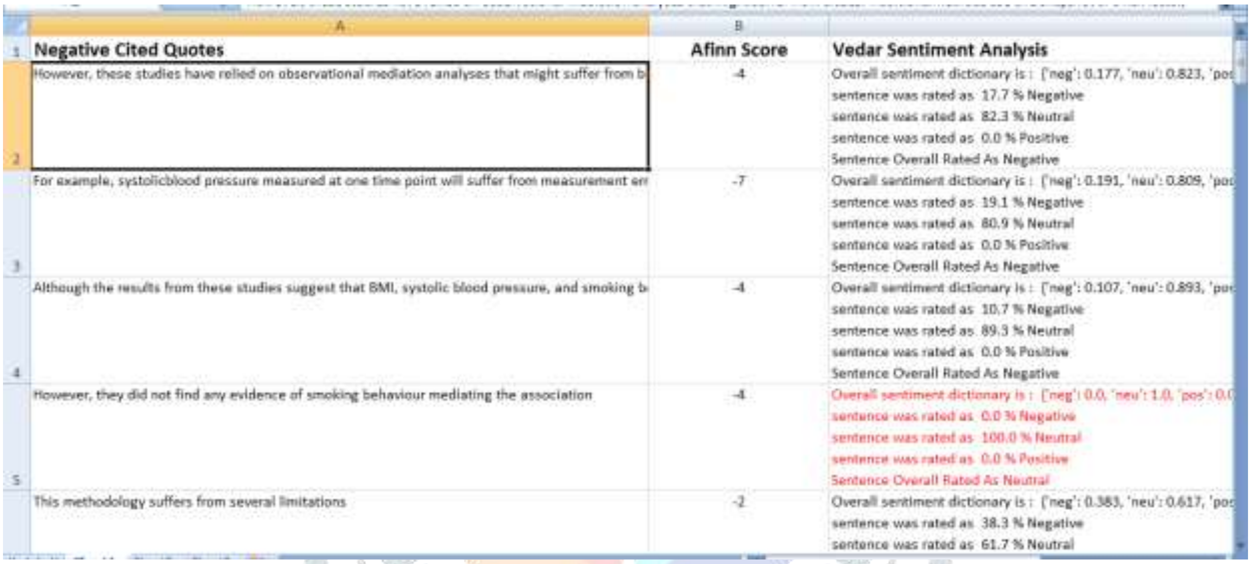
Also impelemented Sentiword Lexicon to identify the similarity of Afinn lexicon negative polarity statements. The essential part of of Sentiword Lexicon is giving results for both polarity along with compound polarity. Though the Sentiword Lexicon specifically for Social media content analysis it has given approximately 70% of similarity when performed with 42 records, only 8 records have been mentioned as dissimilarity. The Figure 3 shows the output of Afinn and Sentiword Lexicons.



```
In [13]: import nltk
         # initialize afinn sentiment analyzer
         from afinn import Afinn
         af = Afinn()
         print(af.score('Another study, which is available in abstract form, failed to show adequate test-retest reliability between o


         print(af.score('This is not surprising because the scales share many items. There have been no studies specifically designed

         1.0
         1.0
```

"Failed to Show " this term is not much express the negativity. Its Showing as "neutral value 0"

Figure: 2 The Low Strength of negative words not determine the negativity of Polarity



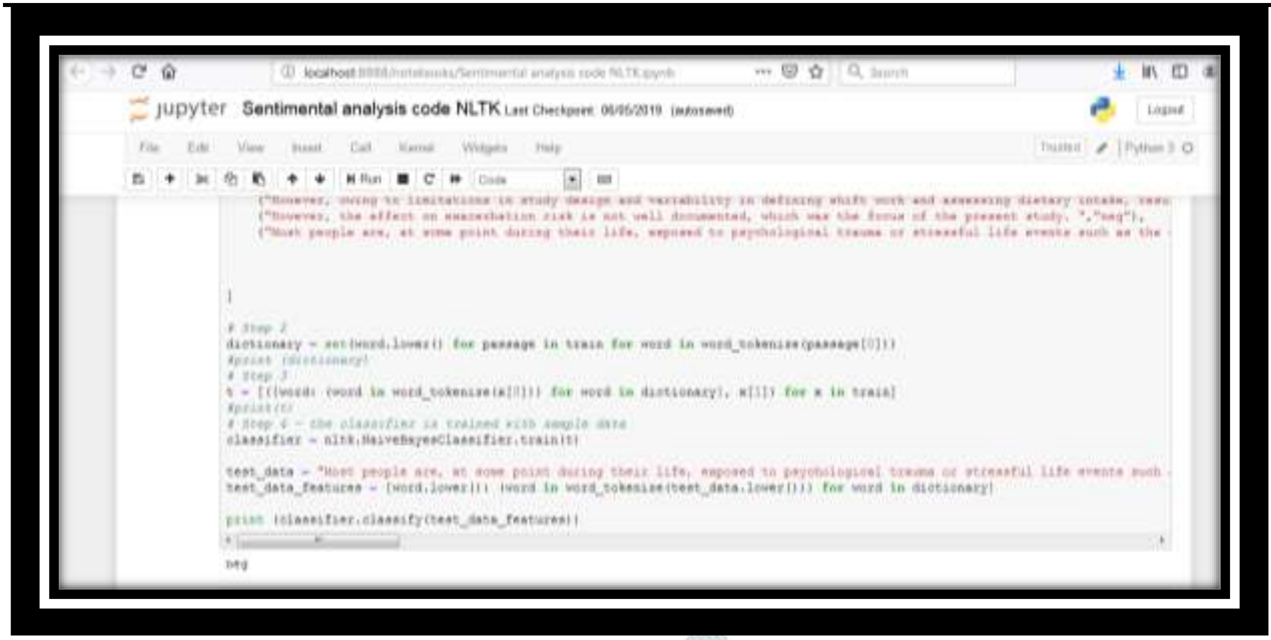
	A	B	
1	Negative Cited Quotes	Afinn Score	Vedar Sentiment Analysis
2	However, these studies have relied on observational mediation analyses that might suffer from b	-4	Overall sentiment dictionary is : {'neg': 0.177, 'neu': 0.823, 'pos sentence was rated as 17.7 % Negative sentence was rated as 82.3 % Neutral sentence was rated as 0.0 % Positive Sentence Overall Rated As Negative
3	For example, systolicblood pressure measured at one time point will suffer from measurement err	-7	Overall sentiment dictionary is : {'neg': 0.191, 'neu': 0.809, 'pos sentence was rated as 19.1 % Negative sentence was rated as 80.9 % Neutral sentence was rated as 0.0 % Positive Sentence Overall Rated As Negative
4	Although the results from these studies suggest that BMI, systolic blood pressure, and smoking b	-4	Overall sentiment dictionary is : {'neg': 0.107, 'neu': 0.893, 'pos sentence was rated as 10.7 % Negative sentence was rated as 89.3 % Neutral sentence was rated as 0.0 % Positive Sentence Overall Rated As Negative
5	However, they did not find any evidence of smoking behaviour mediating the association	-4	Overall sentiment dictionary is : {'neg': 0.0, 'neu': 1.0, 'pos': 0.0 sentence was rated as 0.0 % Negative sentence was rated as 100.0 % Neutral sentence was rated as 0.0 % Positive Sentence Overall Rated As Neutral
	This methodology suffers from several limitations	-2	Overall sentiment dictionary is : {'neg': 0.383, 'neu': 0.617, 'pos sentence was rated as 38.3 % Negative sentence was rated as 61.7 % Neutral

Figure: 3 Results of Afinn score and Sentiword score

3.2 Using NaiveBayseClassifier to Train the Dataset

Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. In the proposed method implemented NaiveBayseClassifier to train dataset for identifying the polarity of new arrival citations. The Limitation of using this algorithm is to find more number of datasets manually and pool into the existing corpus to fine-tune the algorithm in effective way. Below are the steps followed to compute the algorithm.

Step 1: Pooled sample train dataset according to the Afinn Lexicon score value
to mention the range of negative polarity as "negative" or " very negative"
Step 2: Word Tokenizing
Step 3: Trained the existing data using NaiveBayseClassifier
Step 4: Passed new arrival of citation comment to the algorithm for classification
Step 5: Results has been shown according to sentence negative polarity



jupyter Sentimental analysis code NLTK Last Checkpoint: 06/05/2019 (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run Code
Trusted Python 3
Logout

```
    ("However, owing to limitations in study design and variability in defining shift work and assessing dietary intake, rese
    ("However, the effect on exacerbation risk is not well documented, which was the focus of the present study. ","neg"),
    ("Most people are, at some point during their life, exposed to psychological trauma or stressful life events such as the

]

# Step 2
dictionary = set(word.lower() for passage in train for word in word_tokenize(passage[0]))
#print (dictionary)
# Step 3
t = [({word: (word in word_tokenize(x[0])) for word in dictionary}, x[1]) for x in train]
#print(t)
# Step 4 - the classifier is trained with sample data
classifier = nltk.NaiveBayesClassifier.train(t)

test_data = "Most people are, at some point during their life, exposed to psychological trauma or stressful life events such
test_data_features = {word.lower(): (word in word_tokenize(test_data.lower())) for word in dictionary}

print (classifier.classify(test_data_features))
```

```
neg
```

Figure 4: Using of NaiveBayseClassifier to train and classify new arrival of citations

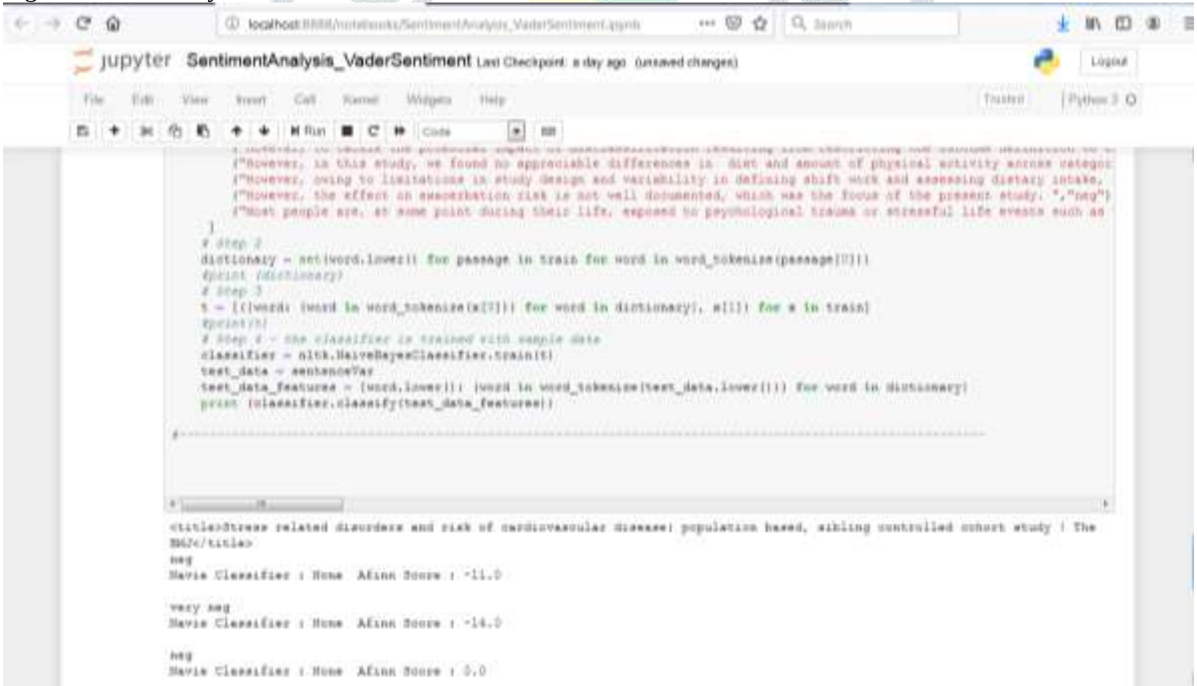
IV. RESULTS AND DISCUSSION

4.1 PYTHON CODE TO STANDARDIZE AUTOMATING OF NEGATIVE CITATION

The reason behind choosing python as language tool to process this research approach is quietly amazing with integrated platform can merge with any sort of languages also equipped with various domain library that give tremendous look and confident to move further to code in it.

The code has been implemented with the following aspects to automate the negative citations of given URL link in dynamic way but it has been designed especially for British medical Journal, later will design a standard format of API for several Journals.(see figure 5)

Step 1: Web scrapping code to extract information from the URL
Step 2: Used regular expression to extract sentences from paragraph
Step 3: Incorporated Afinn Lexicon and Naivye Bayse Classifier algorithm to classify the
polarity of citations
Step 4: Integrated the code by User Defined Functions to Standardize for future enhance



jupyter SentimentAnalysis_VaderSentiment Last Checkpoint: a day ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run Code
Trusted Python 3
Logout

```
    ("However, in this study, we found no appreciable differences in  diet and amount of physical activity across categor
    ("However, owing to limitations in study design and variability in defining shift work and assessing dietary intake,
    ("However, the effect on exacerbation risk is not well documented, which was the focus of the present study. ","neg")
    ("Most people are, at some point during their life, exposed to psychological trauma or stressful life events such as
  ]
# Step 2
dictionary = set(word.lower() for passage in train for word in word_tokenize(passage[0]))
#print (dictionary)
# Step 3
t = [({word: (word in word_tokenize(x[0])) for word in dictionary}, x[1]) for x in train]
#print(t)
# Step 4 - the classifier is trained with sample data
classifier = nltk.NaiveBayesClassifier.train(t)
test_data = sentenceVar
test_data_features = {word.lower(): (word in word_tokenize(test_data.lower())) for word in dictionary}
print (classifier.classify(test_data_features))

#------------------------------------------------------------------------------------------
```

```
<title>Stress related disorders and risk of cardiovascular disease: population based, sibling controlled cohort study | The
BMJ</title>
neg
Navie Classifier : None  Afinn Score : -11.0

very neg
Navie Classifier : None  Afinn Score : -14.0

neg
Navie Classifier : None  Afinn Score : 0.0
```

Figure 5: Output - Standardize code for Automating negative citations

LIMITATION OF THIS STUDY

The proposed method has been implemented a model using python with required alogirthm to find the negative of citations. On future, the research can be enhanced by comparing various classification algorithms to make sure the accuracy of negativity in the citations given by the author. The need of huge data set can reveal the good performance while training the data using classifiers also the architecture can be designed to support multiple journal of clinical trials articles.

CONCLUSION

This study suggested, incorporating Lexicon score value as one of the parameter in classifier to measure the range of negative citation that has been expressed in an article can give much impression of accuracy on deciding the negative polarity. The Implementation of python code can be useful to derive negative polarity for given URL. The challenging part is collecting huge dataset to give much clarity on classify the negative citations. This research gives impact on thriving towards success on assessing qualitative research to analyze negative citations of articles that can help the education field in fruitful way.

REFERENCES

[1] Awais Athar, Sentiment Analysis of Citations using Sentence Structure-Based Features. Proceedings of the ACL-HLT 2011 Student Session, pages 81–87, Portland, OR, USA 19-24 June 2011. c 2011 Association for Computational Linguistics

[2] Christian Catalini, Nicola Lacetera, and Alexander Oettl , The incidence and role of negative citations in science, PNAS , November 10, 2015 , vol. 112 , no. 45 , 13823–13826

[3] Kumar ravi, Vadlamani ravi, Sri rengaraj setlur, Venu Govindaraju, Article Citation Sentiment Analysis Using Deep Learning, Proc. 2018 IEEE 17th Int’l Conf. on Cognitive Informatics & Cognitive Computing (ICCI*CC’18), 978-1-5386-3360-1/18/$31.00 ©2018 IEEE

[4] S. Anupkant, P.V.M. Seravana Kumar, Nayani Sateesh, D. Bhanu Mahesh, Opinion mining on author□s citation characteristics of scientific publications, 978-1-5090-6399-4/17/$31.00_c 2017 IEEE

[5] Diana C. Cavalcanti, Ricardo B. C. Prudêncio, Shreyasee S. Pradhan, Jatin Y. Shah , Ricardo S. Pietrobon, 1082-3409/11 $26.00 © 2011 IEEE, DOI 10.1109/ICTAI.2011.32

[6] G.Parthasarathy, D.C.Tomar, Sentiment Analyzer: Analysis of Journal Citations from Citation Databases, 978-1-4799-4236-7/14/$31.00_c 2014 IEEE

[7] XIAOMEI BAI, IVAN LEE, ZHAOLONG NING, AMR TOLBA, AND FENG XIA, The Role of Positive and Negative Citations in Scientific Evaluation, Received July 3, 2017, accepted August 9, 2017, date of publication August 15, 2017, date of current version September 19, 2017, Digital Object Identifier 10.1109/ACCESS.2017.2740226