International Journal for Innovative Engineering and Management Research

A Peer Revieved Open Access International Journal

www.ijiemr.org

COPY RIGHT

ELSEVIER

SSRN

IJIEMR Transactions, online available on 13th Jul 2022. Link :http://www.ijiemr.org/downloads.php?vol=Volume-11&issue=Issue 07

DOI: 10.48047/IJIEMR/V11/ISSUE 07/03

Title **SENTIMENTAL ANALYSIS OF TWITTER DATA USING NLP AND MACHINE LEARNING TECHNIQUES**

Volume 11, ISSUE 07, Pages: 12-21

Paper Authors

Kotha Jyoshna Priya, Dr.G.Sai Chaitanya Kumar, D.Varun Prasad

SENTIMENTAL ANALYSIS OF TWITTER DATA USING NLP AND MACHINE LEARNING TECHNIQUES

[1] Kotha Jyoshna Priya, [2] Dr.G.Sai Chaitanya Kumar, [3] D.Varun Prasad
[1] PG Student, Department of CSE, DVR&Dr.HS MIC College of Technology, Kanchikacherla,AP
[2] Assoc. Professor, Department of CSE, DVR&Dr.HS MIC College of Technology, Kanchikacherla, AP,
[3]Assoc.Professor, Dept of CSE, DVR&Dr.HS MIC College of Technology, Kanchikacherla, AP
[1] jyoshna.kotta@gmail.com, [2] saichaitanyakumar@mictech.ac.in, [3] varunprasad@mictech.ac.in

Abstract:

Every social networking site, including Facebook, Twitter, Instagram, and others, has emerged as a major information source. It has been discovered that a corporate organisation can benefit from data extraction and analysis from social networking sites for their product promotion. One of the most common platforms for people to voice their opinions and sentiments about a product is Twitter. In our study, we examine how the general public feels about a product using data from Twitter. First, in order to filter tweets, we created a pre-processed data framework based on natural language processing (NLP). In order to analyse sentiment, we also use the Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) model concepts. This project aims to precisely categorise good and negative tweets using BoW and TF-IDF. The accuracy of sentiment analysis can be significantly increased by utilising TF-IDF vectorizer, and simulation results demonstrate the effectiveness of our suggested solution. Using NLP, we were able to analyse sentiment with an accuracy of 85%.

Index Terms—Natural language processing (NLP), Twitter, data mining, Sentiment analysis

Introduction:

The amount of data produced by internet services nowadays is enormous and is growing rapidly every day [1]. Microblogging is practised on social networking sites, where it has developed into a powerful instrument for communication among Internet users [2]–[3].Every business, large and small, has joined social networking sites to share their products and look for customer ratings. In order to measure customer happiness and improve their product, the corporation will employ sentiment analysis to understand how customers feel about their products. In particular, the established approach for sentiment analysis is used frequently to examine relationships between any device, famous person, sports team, and other entities.

After Facebook, Twitter is the second-largest social networking site, producing 21 million tweets each hour and 347,222 tweets per minute [1]. As a result, it opens the door for sentiment analysis and data mining based on user tweets. Since sentiment analyses are a type of data mining, they allow for the observation of consumer sentiment toward a variety of topics and goods. It is also the foundation for techniques like machine learning, computational linguistics, biometrics, and natural language processing. Twitter is our platform of choice for sentiment analysis because it provides options for the sensitivity of articulated disposition. Twitter users can express their concise ideas via a short message because the character limit is 140 on the platform [4].

With the help of the Bag of Word (BoW) model and the TF-IDF (Term Frequency - In- verse Document Frequency) model concept, we have created an NLP-based pre-processed data framework to filter tweets in this paper. Tokenization, stemming, lemmatization, stop word removal, POS tagging, named entity recognition, coreference resolution, and text modelling like Bag of Word and TF IDF Model have all been accomplished using NLP techniques [5].When tweets are gathered using Twitter's streaming API, the major goal is to determine the emotion of each one by defining positive and negative polarity.

These tweets serve as basic data for us. At that moment, we apply the suggested technique that provides a tweet evaluation. Before making a purchase decision, the client will comprehend the feedback on the services from the sentiment analysis. The remainder of the essay is structured as follows. For analysing earlier work, Section II offers similar work. The suggested work's architectural overview and proposed algorithmic rule are briefly detailed in Section III. A basic description of sentiment analysis of a product is given in section IV. The simulated results and performance assessment of our suggested strategy are reported in section V. The paper is concluded in Section VI.

Related Work:

In [6], the authors suggested a machine learning approach for sentiment analysis using an existing twitter dataset. The idea of sentiment analysis utilising a suggested method that categorises Tweets automatically as good, negative, or neutral. Additionally, they are utilising a tree kernel and Part Of Speech (POS)-Specific polarity characteristics. In [7], an ontology-based approach for sentiment analysis is put out.

In order to identify the sentiment behind judgments and provide a description for such polarisation, the authors devised a model that combines domain ontology with natural language processing techniques. The technique tests were created employing the two distinct mediums of movies and digital cameras.[8].The techniques of sentiment analysis are trained in [9]–[10] to identify sentiment polarity, which allows them to automatically track down sentiments from various documents, blogs, sentences, or words. Through the integration of semantic technology, NLP, and information extraction, the authors of [11] created a new approach for semantic knowledge extraction from research materials and articles.

Using data cleansing, data extraction, and data consolidation, the authors of [12] proposed a novel method for extracting structured data from emails. The supervised learning approach, according to the authors' theory in [13], is built on label datasets that are trained to provide useful outputs. Applying the Naive Bayes method, maximum entropy, and support vector machines to monitor the learning process enables successful sentiment analysis. In [14], the authors demonstrated that they could achieve an accuracy of up to 82.1 percent using the Naive Bayes algorithm. The accuracy of the sentiment analysis performed by the authors in [15] using the K-Nearest Neighbor classifier was 74.74 percent. The authors of [16] presented a survey on sentiment analysis of Twitter data using several methodologies. Different machine learning techniques, including Naive Bayes, Maximum Entropy, and Vector Machine Support, were employed to analyse sentiment and show the efficacy of various feature sizes.

Proposed Work

This section summarises our ideas and focuses on a method for conducting sentiment analysis on Twitter data. Figure 1 depicts the architectural overview defining a general process design for sentiment analysis. The created technique is based on three crucial components: Data Extraction from a specific project or product, Natural Language Toolkit (NLTK) preprocessing of the extracted Tweets, and Classifier model that determines the sentiment of each Tweet.

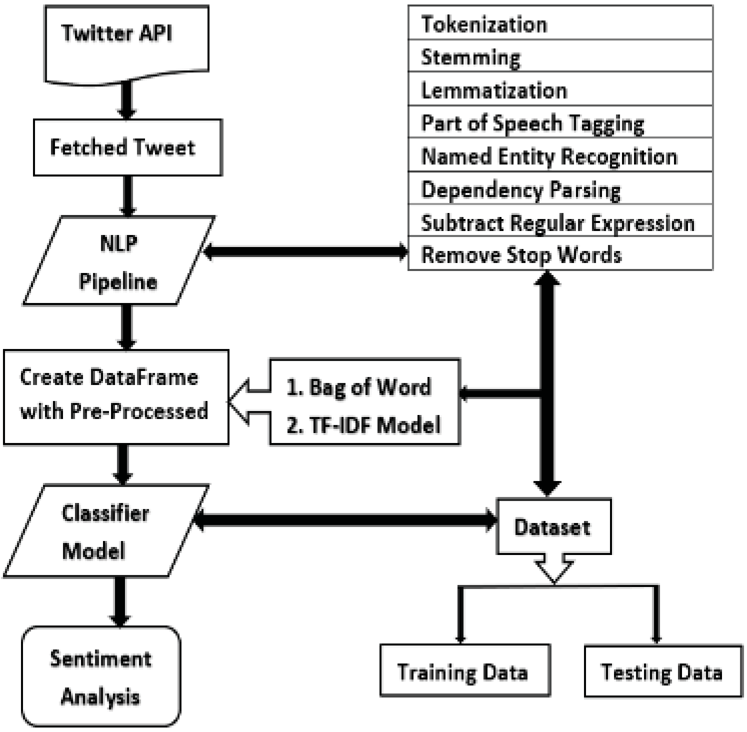


Fig. 1: Architectural overview of proposed work

Data extraction, tokenization, stemming, lemmatization, stopword removal, parts of speech tagging, named entity recognition, create a data frame, text modelling [17], and a classifier model are all used to perform sentiment analysis on the Twitter data. Each of these processes has its own algorithm and set of packages that must be processed. The TF-IDF model has been used to extract crucial terms from tweets in order to forecast sentiment. Pickle module, which has already completed the object serialisation process, aids in the construction of a classifier model.

In order to find the attitudes, we first constructed a classifier model using a dataset [18] that is displayed in Algorithm 1.Additionally, we created a sentiment analysis and tweet filtering pre-processed data framework based on natural language processing (NLP), as demonstrated in 2.In essence, we used the BoW model to classify documents and model their text before feeding them into our method for analysing tweets. BoW does not preserve any semantic information and accords all words the same significance. In order to identify the most crucial word in tweets, we override BoW and include the TF-IDF model. And it aids in highly accurate sentiment analysis. Our suggested system determines whether a tweet is favourable or negative based on the classifier model.

Sentiment Analysis

This study presents the sentiment analysis approach we created for Twitter using Twitter data. We use the comparison between the two most popular devices, the iPhone and the Samsung, as an example. On an iPhone, we downloaded 100 tweets from Twitter, which are displayed. We have deleted stop words and superfluous information from tweets before applying sentiment analysis, which was covered in the previous section, using an NLP pipeline. We now employed "vectorizer" and "classifier," which are "clf" objects built from "tfidfmodel.pickle" and "classifier.pickle," to forecast sentiment. We display the sentiment forecast for each tweet for the iPhone device.

To test if our classifier model correctly predicts the sentiment of a tweet like the one in, we now take a sample of a tweet. Since we do not analyse tweets based on emoticons, an emoticon was used in this tweet. The NLP pipeline removed the emoticon and many redundant words from the tweet, such as "https://t.co/c1yNjARvif," which was changed to "ynjarvif" by using the Python re package to preprocess the tweet. Since this word has no meaning, the TF-IDF model would return zero. Given 'zero' polarity, the classifier model in this instance predicts the sentence to be a negative statement.

It merely makes a deduction regarding the sentiment analysis. For better comprehension and visualisation, we have displayed several sentiments based on the amount of tweets collected, such as 50, 100, 500, and 1000, respectively. Within a year, two businesses released a large number of phones, some of which gained greater notoriety than those from competing companies. Therefore, it is impossible to state one company is more well-known than another because the difference in the proportion of favourable to negative tweets between the two phones isn't substantial. However, we can argue that both of these phone companies are more well-known than other phone companies.

Natural Language Processing (NLP)

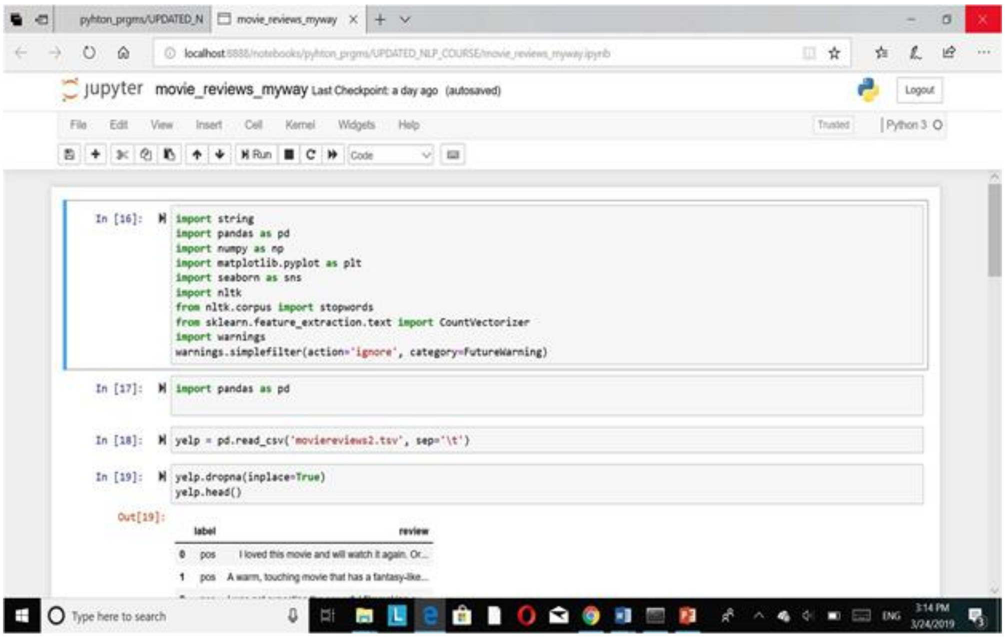
Natural Language Processing (NLP) is a cornerstone of modern AI and can be used to enable any number of different business use cases, from speech recognition to chatbots and entiment analysis. NLP can be used to help AI applications better understand what a given user wants and then as part of a larger platform, actually help the user to execute whatever action or operation they need. Jason Flaks understands what NLP is capable of more than most: He was the leading inventor of the conversational technology behind Microsoft’s Kinect and Hololens.

Now, Flaks is the CTO of startup Xembly, which is officially launching today. Flaks leads Xembly alongside CEO Pete Christothoulou, with the goal to use NLP and conversational AI to enable what the company refers to as an “AI chief of staff.”

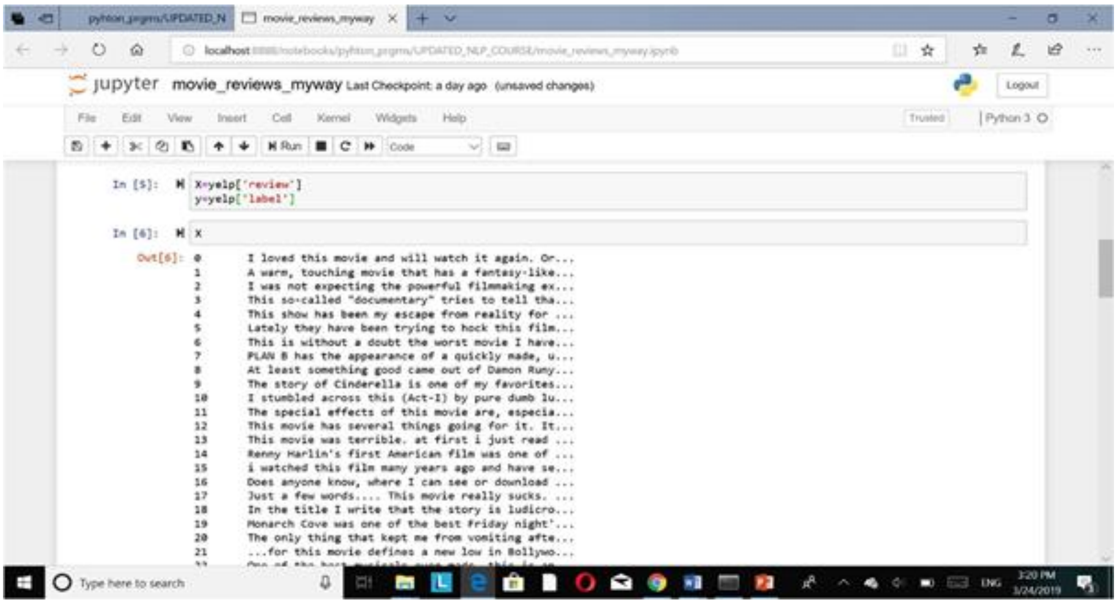
Helping business people and individuals to better organize and schedule their work lives is a complicated task. A top level business executive might be lucky enough to have a chief of staff to help schedule meetings, take notes and determine action items for follow up. Xembly aims to fill a similar role, with an automated NLP-powered AI platform. Modern conversational AI technologies and NLP can listen into a roomful of people and understand when more than just one person is talking and who is talking to whom. In Flaks’ view, modern conversational AI technologies can actually be an active listener, even in a room full of people, rather than just waiting for a single voice to request a single action. The real magic is, ultimately, what you can do with that,” Flaks said. “I think that’s where we continue to want to expand. The question is, how do we take that data and do meaningful things with it?”

The ability to transcribe notes is not a new thing for NLP. What Xembly is doing is going beyond just spoken dialogue to producing well-written prose for action items and meeting summaries. Flaks said that it takes a really complex set of machine learning models to be able to achieve that. He explained that Xembly runs its own machine learning models to detect action items from a conversation or interaction with Xena. The action items can include items like pulling out due dates and order requests. The Xembly system is also able to conduct topic segmentation for conversations to help users better understand the flow of a meeting.

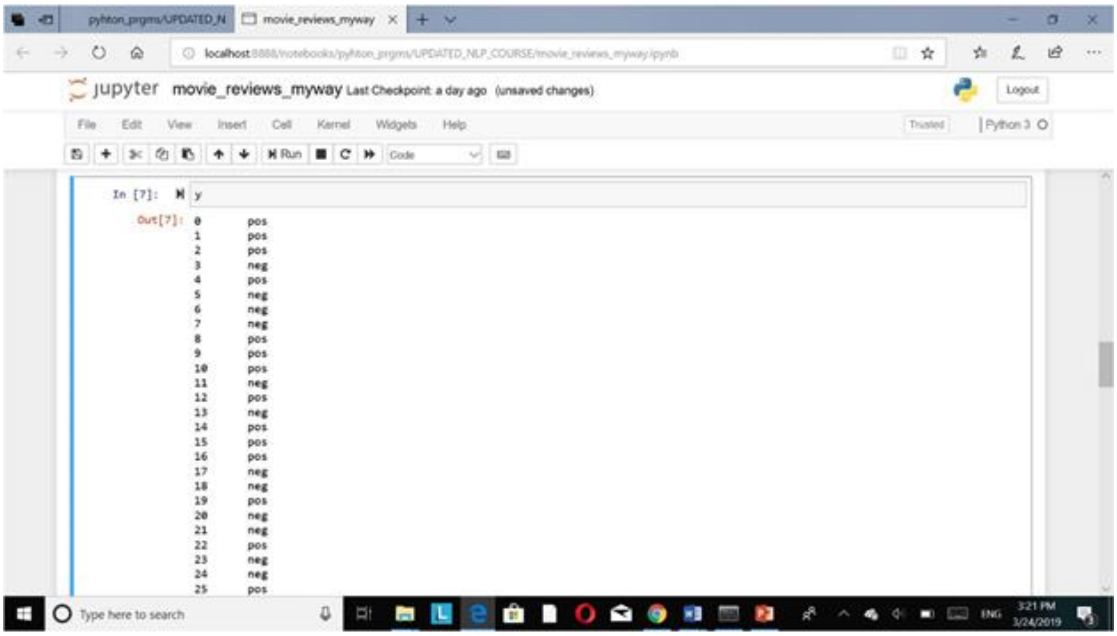
Results and Outputs


pyhton_prgms/UPDATED_N
movie_reviews_myway
localhost:8888/notebooks/pyhton_prgms/UPDATED_NLP_COURSE/movie_reviews_myway.ipynb
jupyter movie_reviews_myway Last Checkpoint: a day ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
ENG
3:14 PM
3/24/2019

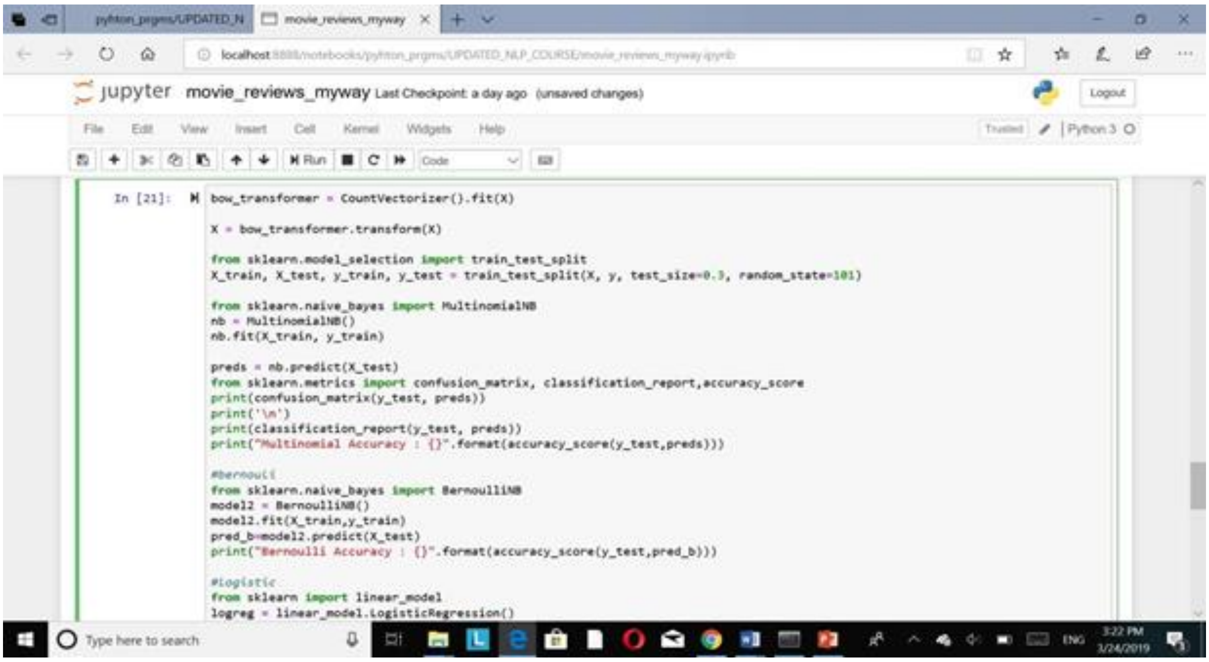
```
In [16]:  import string
          import pandas as pd
          import numpy as np
          import matplotlib.pyplot as plt
          import seaborn as sns
          import nltk
          from nltk.corpus import stopwords
          from sklearn.feature_extraction.text import CountVectorizer
          import warnings
          warnings.simplefilter(action='ignore', category=FutureWarning)

In [17]:  import pandas as pd

In [18]:  yelp = pd.read_csv('moviereviews2.tsv', sep='\t')

In [19]:  yelp.dropna(inplace=True)
          yelp.head()

Out[19]:
     label                                    review
  0  pos      I loved this movie and will watch it again. Or...
  1  pos   A warm, touching movie that has a fantasy-like...
```


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

```
In [5]:  X=yelp['review']
         y=yelp['label']

In [6]:  X

Out[6]: 0         I loved this movie and will watch it again. Or...
        1         A warm, touching movie that has a fantasy-like...
        2         I was not expecting the powerful filmmaking ex...
        3         This so-called "documentary" tries to tell tha...
        4         This show has been my escape from reality for ...
        5         Lately they have been trying to hock this film...
        6         This is without a doubt the worst movie I have...
        7         PLAN B has the appearance of a quickly made, u...
        8         At least something good came out of Damon Runy...
        9         The story of Cinderella is one of my favorites...
        10        I stumbled across this (Act-I) by pure dumb lu...
        11        The special effects of this movie are, especia...
        12        This movie has several things going for it. It...
        13        This movie was terrible. at first i just read ...
        14        Renny Harlin's first American film was one of ...
        15        i watched this film many years ago and have se...
        16        Does anyone know, where I can see or download ...
        17        Just a few words.... This movie really sucks. ...
        18        In the title I write that the story is ludicro...
        19        Monarch Cove was one of the best Friday night'...
        20        The only thing that kept me from vomiting afte...
        21        ...for this movie defines a new low in Bollywo...
```

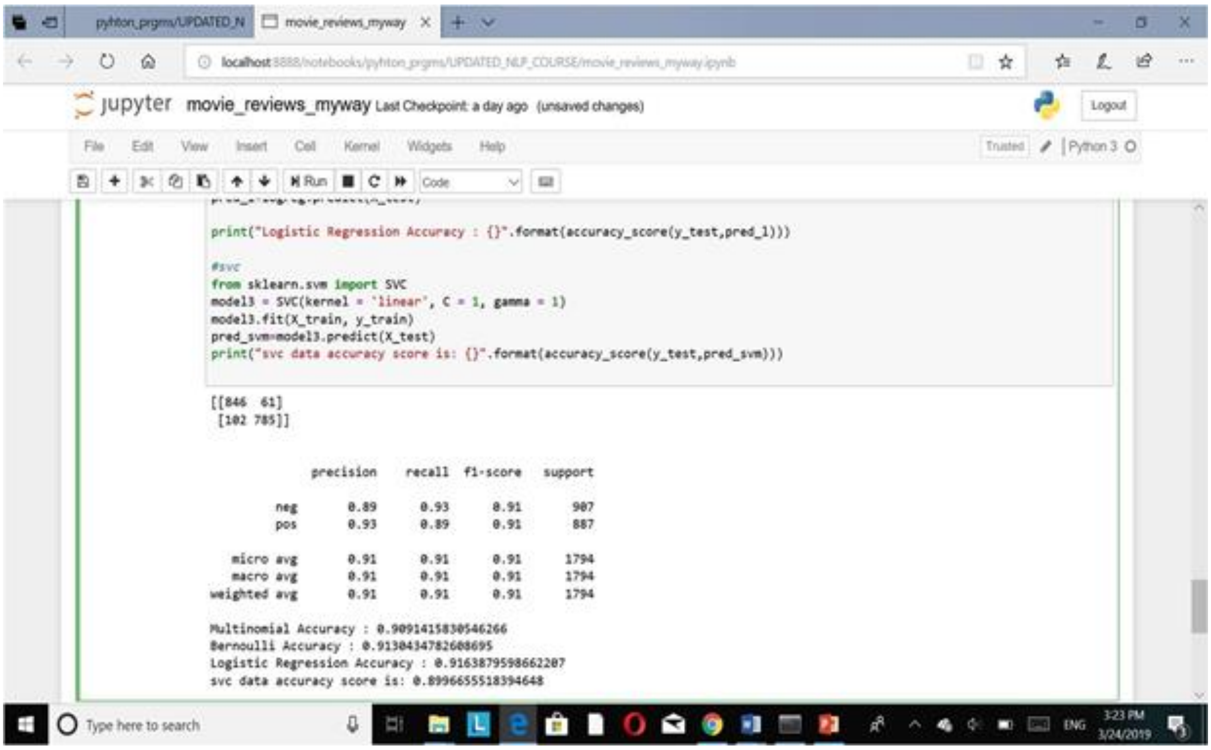

jupyter movie_reviews_myway Last Checkpoint: a day ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Logout
Run
Code
Type here to search

```
In [7]:  y

Out[7]: 0         pos
        1         pos
        2         pos
        3         neg
        4         pos
        5         neg
        6         neg
        7         neg
        8         pos
        9         pos
        10        pos
        11        neg
        12        pos
        13        neg
        14        pos
        15        pos
        16        pos
        17        neg
        18        neg
        19        pos
        20        neg
        21        neg
        22        pos
        23        neg
        24        neg
        25        pos
```



jupyter movie_reviews_myway Last Checkpoint: a day ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Trusted
Python 3
Logout
Type here to search
3:22 PM
3/24/2019

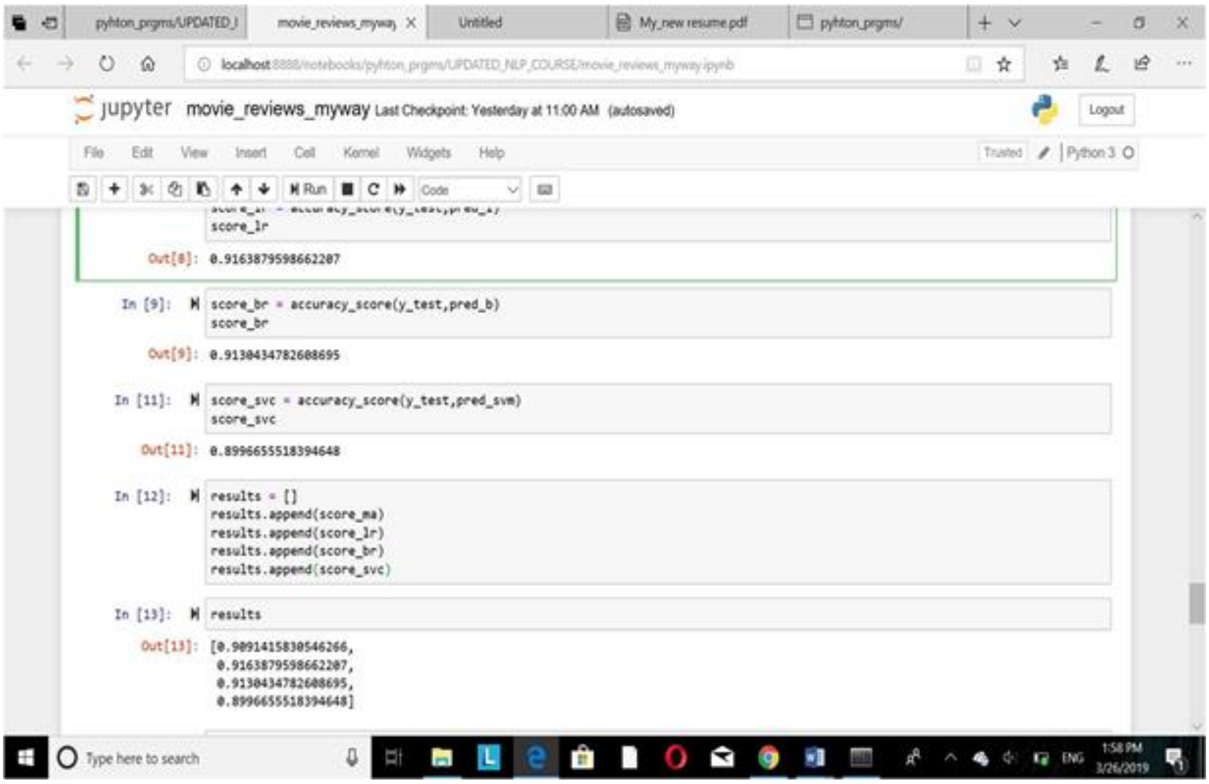
```
In [21]:  bow_transformer = CountVectorizer().fit(X)

          X = bow_transformer.transform(X)

          from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)

          from sklearn.naive_bayes import MultinomialNB
          nb = MultinomialNB()
          nb.fit(X_train, y_train)

          preds = nb.predict(X_test)
          from sklearn.metrics import confusion_matrix, classification_report,accuracy_score
          print(confusion_matrix(y_test, preds))
          print('\n')
          print(classification_report(y_test, preds))
          print("Multinomial Accuracy : {}".format(accuracy_score(y_test,preds)))

          #bernoulli
          from sklearn.naive_bayes import BernoulliNB
          model2 = BernoulliNB()
          model2.fit(X_train,y_train)
          pred_b=model2.predict(X_test)
          print("Bernoulli Accuracy : {}".format(accuracy_score(y_test,pred_b)))

          #logistic
          from sklearn import linear_model
          logreg = linear_model.LogisticRegression()
```



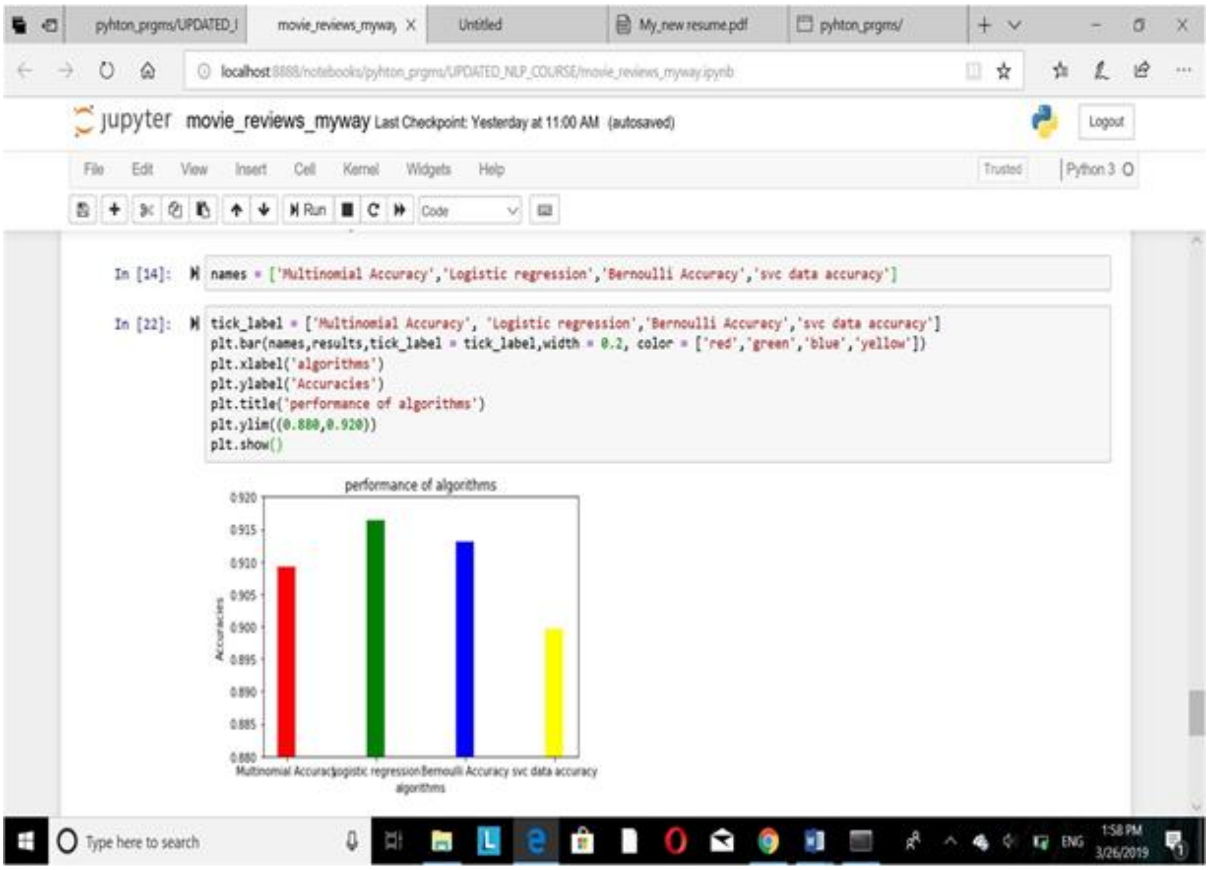
jupyter movie_reviews_myway Last Checkpoint: a day ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
Trusted
Python 3
Logout
Type here to search
3:23 PM
3/24/2019

```
print("Logistic Regression Accuracy : {}".format(accuracy_score(y_test,pred_l)))

#svc
from sklearn.svm import SVC
model3 = SVC(kernel = 'linear', C = 1, gamma = 1)
model3.fit(X_train, y_train)
pred_svm=model3.predict(X_test)
print("svc data accuracy score is: {}".format(accuracy_score(y_test,pred_svm)))
```

```
[[846  61]
 [102 785]]


              precision    recall  f1-score   support

         neg       0.89      0.93      0.91       907
         pos       0.93      0.89      0.91       887

   micro avg       0.91      0.91      0.91      1794
   macro avg       0.91      0.91      0.91      1794
weighted avg       0.91      0.91      0.91      1794

Multinomial Accuracy : 0.9091415830546266
Bernoulli Accuracy : 0.9130434782608695
Logistic Regression Accuracy : 0.9163879598662207
svc data accuracy score is: 0.8996655518394648
```


pyhton_prgms/UPDATED_I
movie_reviews_myway
Untitled
My_new resume.pdf
pyhton_prgms/
localhost:8888/notebooks/pyhton_prgms/UPDATED_NLP_COURSE/movie_reviews_myway.ipynb
jupyter movie_reviews_myway Last Checkpoint: Yesterday at 11:00 AM (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
1:58 PM
3/26/2019

```
          score_lr
 Out[8]:  0.9163879598662207

  In [9]: score_br = accuracy_score(y_test,pred_b)
          score_br
 Out[9]:  0.9130434782608695

 In [11]: score_svc = accuracy_score(y_test,pred_svm)
          score_svc
Out[11]:  0.8996655518394648

 In [12]: results = []
          results.append(score_ma)
          results.append(score_lr)
          results.append(score_br)
          results.append(score_svc)

 In [13]: results
Out[13]:  [0.9091415830546266,
           0.9163879598662207,
           0.9130434782608695,
           0.8996655518394648]
```



```
In [14]: names = ['Multinomial Accuracy','Logistic regression','Bernoulli Accuracy','svc data accuracy']

In [22]: tick_label = ['Multinomial Accuracy', 'Logistic regression','Bernoulli Accuracy','svc data accuracy']
         plt.bar(names,results,tick_label = tick_label,width = 0.2, color = ['red','green','blue','yellow'])
         plt.xlabel('algorithms')
         plt.ylabel('Accuracies')
         plt.title('performance of algorithms')
         plt.ylim((0.880,0.920))
         plt.show()
```

Conclusion:

In this paper, a basic yet novel methodology on sentimental analysis of motion picture surveys is performed utilizing promising supervised AI calculations. The out comes obtained concludes Logistic Regression as the best classifier among others in accomplishing 92% exactness for substantial number of movie reviews. Infuture, we attempt to explore its effectiveness considering enormous data set sutilizing the unsupervised and semi supervised AI methods. The NLP techniques for text pre processing are well performed to obtained data to for building the model. These techniques are far better than any other for the data preprocessing as it states that text classification is the main step forth emining of the data for the sentimental analysis. A conspicuous method to stretch out this work is add other order calculations to the, e.g., Conditional Random Fields or progressively expound gatherings. There are additionally a few highlights and highlight choice techniques that could be explored and a less innocent method for dealing with nullification. Instead of the straightforward treatment of in validation utilized here, away to deal with programmed enlistment of degree through a nullification finder could be utilized. The Future Sentimental Analysis work can be done on the group of the images to identify the emotions of the faces in the images and make the best filter upgrade for the group of images by identifying the good emotions the work can performed by the unsupervised algorithms

Bibliography

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval,vol.2,pp.1- 135,2008.

[2] A.BifetandE.Frank,"Sentimentknowledgediscoveryintwitterstreamingdata,"in DiscoveryScience,2010,pp.1-15.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machinelearning techniques," in Proceedings of the ACL-02 conference on Empirical methods in naturallanguageprocessing-Volume10,2002,pp.79-86.

[4] R.Basili,A.Moschitti,andM.T.Pazienza,"Languagesensitivetextclassification,"inRIAO,2000,pp.331-343.

[5] P. S. Jacobs, "Joining statistics with NLP for text categorization," in Proceedings of the thirdconferenceonAppliednaturallanguage processing,1992,pp.178-185.

[6] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization,"Machinelearning,vol.39,pp.135-168,2000.

[7] V.N.VapnikandV.Vapnik,Statisticallearningtheoryvol.1:WileyNewYork,1998.

[8] G.SaltonandM.J.McGill,"Introductiontomoderninformationretrieval,"1986.

[9] S. Dumais, J. Platt, D. Heckerman, andM. Sahami, "Inductive learning algorithms andrepresentations for text categorization," in Proceedings of the seventh international conference onInformationandknowledgemanagement, 1998,pp.148-155.

[10] S.M.Weiss,C.Apte,F.J.Damerau,D.E.Johnson,F.J.Oles,T.Goetz,etal.,"Maximizingtext-miningperformance,"IEEEIntelligentsystems,pp.63-69,1999.

[11] R. Feldman, "Techniques and

applications for sentiment analysis," Communications of theACM, vol. 56, pp. 82-89, 2013.

[12] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet3.0:AnEnhancedLexicalResourceforSentimentAnalysisandOpinionMining,"inLREC,2010,pp.2200-2204.

[13] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining,"inLREC,2010,pp.1320-1326.

[14] T.Blog, "Insightsintothe # World Cup conversation on Twitter," in TwitterBlog,ed, 2014.

[15] D. Terrana, A. Augello, and G. Pilato, "Automatic Unsupervised Polarity Detection on aTwitter Data Stream," in Semantic Computing (ICSC), 2014 IEEE International Conference on,2014,pp.128-134.

[16] L.Zhang,"SentimentanalysisonTwitterwithstockpriceandsignificantkeywordcorrelation,"2013.

[17] A.G.Jivani,"Acomparativestudyofstemmingalgorithms,"Int.J.Comp.Tech.Appl,vol.2,pp.1930-1938,2011. . Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA datamining software: anupdate, " ACMSIGKDD explorationsnewsletter,vol.11,pp.10-18,2009.

[18] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization,"Technometrics,vol.49,pp. 291-304,2007.

[19] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," vol. 198, p.282,2004

[20] G.Sai Chaitanya Kumar, Dr.Reddi Kiran Kumar, Dr.G.Apparao Naidu, “Noise Removal in Microarray Images using Variational Mode Decomposition Technique” Telecommunication computing Electronics and Control ISSN 1693-6930 Volume 15, Number 4 (2017), pp. 1750-1756

[21] S. Gorintla, B. A. Kumar, B. S. Chanadana, N. R. Sai and G. S. C. Kumar, "Deep-Learning-Based Intelligent Pothole Eye+ Detection Pavement Distress Detection System," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2022, pp. 1864-1869, doi: 10.1109 / ICAAIC53929.2022.9792696.