VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI-590018

INTERNSHIP REPORT

ON

"INFORMATION EXTRACTION USING TWITTER COMMENTS"

Submitted in partial fulfillment of the requirements for the award of

Degree of Bachelor of Engineering in computer science and technology

Submitted by

AsmithaThulapule	**4SU16CS018**
Monisha B L	**4SU17CS047**
Rahul T C	**4SU17CS070**
ShreedharGouli	**4SU17CS122**

Under the Guidance of

Sheik imam

Carried out at

EBRAIN SOFT PVT.LTD

Bengaluru

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SDM INSTITUTE OF TECHNOLOGY

UJIRE-574240

2020-2021

S. D. M. INSTITUTE OF TECHNOLOGY

(Affiliated To Visvesvaraya Technological University, Belagavi)

UJIRE-574240

Department of Computer Science and Engineering

CERTIFICATE

This is to Certify that the project work entitled "**INFORMATION EXTRACTION USING TWITTER COMMENTS**" is carried out by ASMITHA THULAPULEbearing USN **4SU16CS018**, MONISHA B Lbearing USN **4SU17CS047**, RAHUL T C bearing USN **4SU17CS070 and** SHREEDHAR GOULIbearing USN **4SU17CS122** in partial fulfilment for requirements for **6TH SEMESTER** of the Visvesvaraya Technological University, Belagavi duringthe year 2020-21. It is certified that all corrections / suggestions indicated for Internship have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the eBrain internship.

Signature of the Guide	**Signature of the H.O.D**
(Guide NAME)	**(Dr.Thayagaraju G. S)**

ABSTRACT

Text mining is a new and exciting research area that tries to solve the information overload problem byusing techniques from machine learning, natural language processing (NLP), data mining, informationretrieval (IR), and knowledge management.

Text mining involves the pre-processing of document collections such as information extraction, term extraction, text categorization, and storage ofintermediate representations.

The techniques that are used to analyse these intermediate representationssuch as clustering, distribution analysis, association rules and visualisation of the results.

The field of text mining seeks to extract useful informationfrom unstructured textual data through the identification and explorationof interesting patterns.

The techniques employed usually do not involvedeep linguistic analysis or parsing, but rely on simple “bag-of-words” textrepresentations based on vector space.

Several approaches to the identification of patterns are discussed, including dimensionality reduction,automated classification and clustering.

ACKNOWLEDGEMENT

The internship opportunity we had with EBRAIN was a great chance for learning and professional development.Therefore, we consider ourselves as a lucky and we were provided to be a part of it.

We would like to thank SHEIK IMAM (ebrain Ltd) for providing us the opportunity for this training.

People working in the workspace made the training huge success and pleasant experience.

We extend our warm gratitude and regards to everyone who helped us during our internship.

TABLE OF CONTENTS

ch.no	Particulars	Page no
1	Company profile	1-3
2	Basic study	4-7
3	Project introduction	8
4	Problem definition	9
5	Problem solution	10-11
6	Screenshots	12-15
7	Results	16
8	Conclusion	17
9	References	18

E-BRAIN SOFTECH

E-BRAIN Softech Pvt. Ltd. is a leading software development firm and training body, has been operational in Karnataka with a dedicated panel of experts from IT Industry. We provide services on Management, Education technology and guidance to anyone looking on any areas of interest. We are a team of qualified, experienced trainers & IT Professionals motivated to educate people by training & nurturing them to the best of their strengths.

1.1 Services

Software Development

Websites, apps, interfaces, AI, and more — we are a full-service product strategy, design, and development partner.

User Experience

Focusing on the customer, we design your customer's journey, optimizing every touchpoint for convenience and delight.

Training

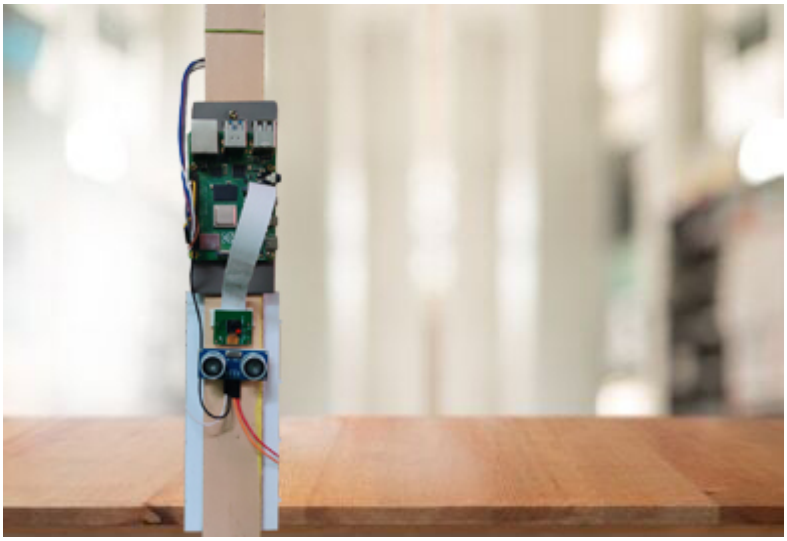
Machine Learning, Deep Learning, Python, Angular, Node JS, Core/Advance Java, Image Processing, Matlab, C/C++, Networking, Android

1.2 products

Digital Eye

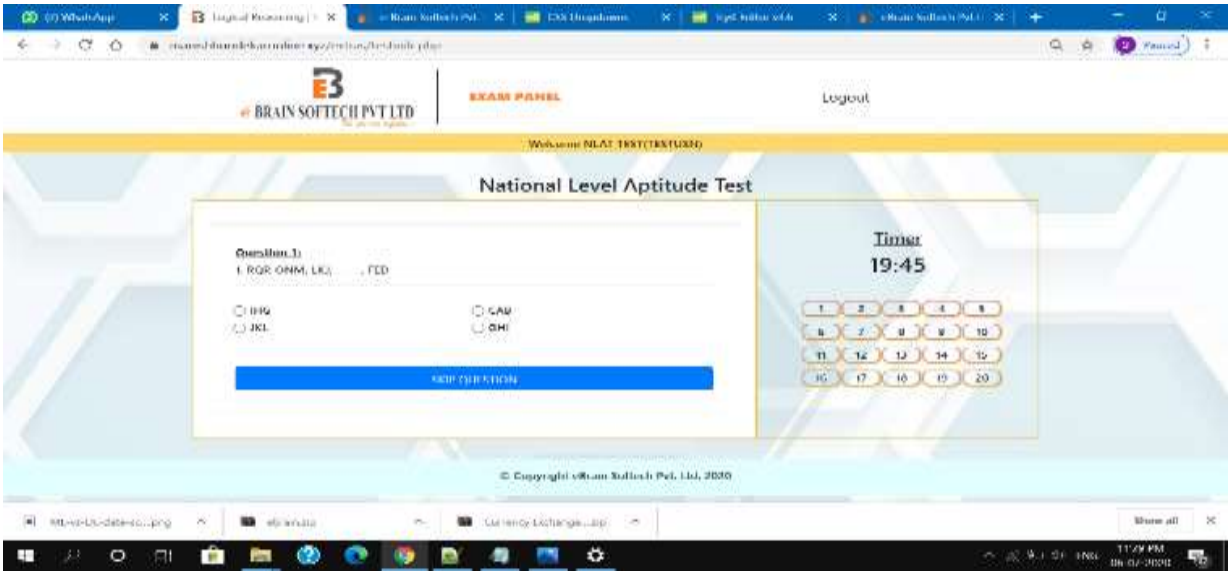
According to the recent survey, India is now home to the largest number of blind people. Forty per cent of the world's blind are in India and the official number is 12 million. The problems of blind and visually impaired people in India range from lack of basic necessities to global issues and prevailing social stigma attached to it. In India, the major issues faced by the blind and visually impaired persons are lack of disabled friendly infrastructure and transportation facilities as well. Most blind girls are not allowed to get out of the house for years together. Nevertheless, the blind deserve the same quality of life as that of a sighted person and have the right to participate equally in the society. As a step towards this; smart canes/sticks can prove to be extremely useful for independent and safe navigation of the blinds.

E-Brain employs a resourceful effort in initiating Digital Eye, which envisions easing the pain of people with visual impairment and blind to be self-relined. Digital Eye is a smart navigating-stick with ultrasonic proximity sensors to detect the presence of target objects and GPS module to guide these blind people to reach out their destinations independently. This stick comprises camera equipped with the object detection algorithm integrated with ultrasonic sensors to detect any upcoming obstacles and to sense the exact distance from that obstacle using voice based bot and a water-detection model integrated with the voicebot to discern between water and no-water regions.



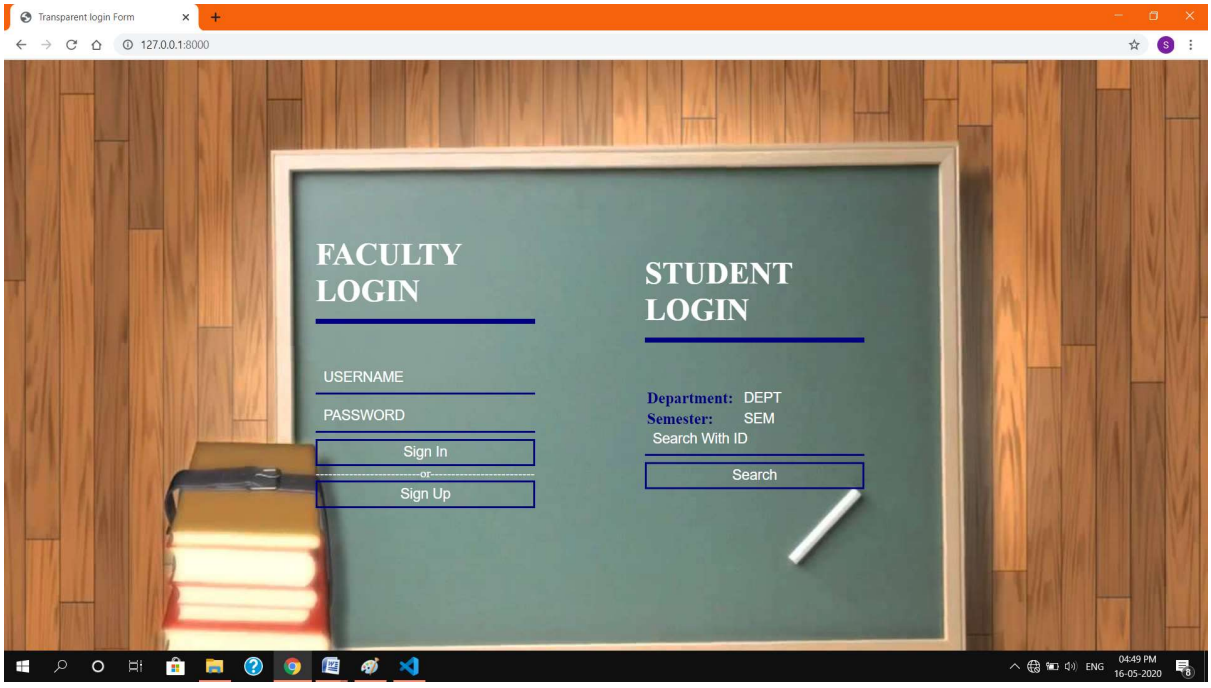
Onlinex

This software provides a best platform for MCQ Examinations and auto evaluation. This software also provides easy evaluation of scores automatically. After Successful Evaluation it will generate a score card for every student Individually. These Report cards will be sent to their Registered Email Id automatically without any man power.



AI Attendance

The proposed software takes the participation of the student using facial identification technique. This participation is recorded by a high quality camera which is installed as a part of classroom and continuously capturing the images of the students, detects their faces, contrast distinguished appearances and mark the attendance. After Marking the attendance, it will also generate a report and will be stored for further use. With that the software also provides an acknowledgment message to the student if he is absent.



BASIC STUDY

2.1 PYTHON

Python is an interpreted, object-oriented, high level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including procedural object-oriented, and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library

The Python 2 language, i.e. Python 2.7.x, was officially discontinued on 1 January 2020 (first planned for 2015) after which security patches and other improvements will not be released for it. With Python 2's end-of-life, only Python 3.5.x and later are supported.

Python uses dynamic typing and a combination of reference counting and a cycle-detecting garbage collector for memory management. It also features dynamic name resolution (late binding), which binds method and variable names during program execution.

2.2 ANACONDA NAVIGATOR

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage con packages, environments, and channels without using command-line commands.

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages, and update them – all inside Navigator.

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system conda. The Anaconda distribution includes data-science packages suitable for Windows, Linux, and MacOS.

Anaconda distribution comes with 1,500 packages selected from PyPI as well as the conda package and virtual environment manager. It also includes a GUI, **Anaconda Navigator**, as a graphical alternative to the command line interface (CLI).

The big difference between conda and the pip package manager is in how package dependencies are managed, which is a significant challenge for Python data science and the reason conda exists.

many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

2.3 JUPYTER NOTEBOOK

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

A Jupyter Notebook can be converted to a number of open standard output formats

Jupyter Notebook can connect to many kernels to allow programming in many languages. By default,Jupyter Notebook ships with the IPython kernel. As of the (October 2014), there are currently 49 Jupyter-compatible kernels for many programming languages, including Python, R, Julia and Haskell.

In addition to running your code, it stores code and output, together with markdown notes, in an editable document called a notebook. When you save it, this is sent from your browser to the notebook server, which saves it on disk as a JSON file with extension.

2.4 PANDAS

Pandas is a library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series, which is a Panel Data.

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the python programming language.

2.5 NUMPY

NumPy is a package in Python used for Scientific Computing. NumPy package is used to perform different operations. The ndarray (NumPy Array) is a multidimensional array used to store values of same datatype. In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

NumPy targets the CPython reference implementation of Python, which is a nonoptimizing bytecode interpreter. Mathematical algorithms written for this version of Python often run much slower than compiled equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays, requiring rewriting some code, mostly inner loops using NumPy.

2.7 MATPLOTLIB

Several toolkits are available which extend Matplotlib functionality. Some are separate downloads, others ship with the Matplotlib source code but have external dependencies.

 Basemap: map plotting with various map projections, coastlines, and political boundaries.

 Cartopy: a mapping library featuring object-oriented map projection definitions, and arbitrary point, line, polygon and image transformation capabilities. (Matplotlib v1.2 and above)

 Excel tools: utilities for exchanging data with Microsoft Excel

 GTK tools: interface to the GTK+ library

 Qt interface

 Mplot3d: 3-D plots

 Natgrid: interface to the natgrid library for gridding irregularly spaced data.

2.8 MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a field of study within machine learning, and focuses on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

There are also some types of machine learning algorithm that are used in very specific case,but three main methods areused today.

i.supervised learning:supervised learning is one of the most basic types of machine laearning.in this type, the machine learning algorithm is trained on labelled data. Even though the data needs to be labelled accurately for this method to work, supervised learning is extremely power full when used in the right circumstances.

ii. unsupervised learning: unsupervised machine learning holds the advantages of being able to work with unlabelled data. this means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program.

iii.reinforcement learning:this learning directly takes inspiration from how human beings learn from data in their lives.it features an algorithm that improves upon itself and learns fromnew situation using a trail-and-error method. favourable outputs area encouraged or ‘reinforced’, and non-favourable outputs are discouraged.

PROJECT INTRODUCTION

Text mining, also known as text analysis, is the process of transforming unstructured text data into meaningful and actionable information. Thisproject utilizes different ML technologies to automatically process data and generate valuable insights, enabling companies to make data-driven decisions.

For businesses, the large amount of data generated every day represents both an opportunity and a challenge. On the one side, data helps companies get smart insights on people's opinions about a product or service. Think about all the potential ideas that you could get from analysing emails, product reviews, social media posts, customer feedback, support tickets, etc.

On the other side, there's the dilemma of how to process all this data. And that's where text mining plays a major role.

Like most things related to Natural Language Processing (NLP), text mining may sound like a hard-to-grasp concept. But the truth is, it doesn't need to be. This guide will go through the basics of text mining, explain its different methods and techniques, and make it simple to understand how it works. You will also learn about the main applications of text mining and how companies can use it to automate many of their processes:

The Natural language texts have information, which is not suitable for computers for analysis purpose. Where as computers uses large amount of text and extract useful information from passages, phrases or single words. So Information Extraction can be considered as restricted form of natural language understanding and here we know about the semantic information, we are seeking for. The task of information Extraction is to extract parts of text and assign specific attribute to it.

Thanks to text mining, businesses are being able to analyse complex and large sets of data in a simple, fast and effective way. At the same time, companies are taking advantage of this powerful tool to reduce some of their manual and repetitive tasks, saving their teams precious time and allowing customer support agents to focus on what they do best.

PROJECT DEFINATION

Information Extraction is one of a number of emerging technologies thatenable the presentation information developed from text in a variety of graphicformats.

This project can be useful for work teams is by providing smart insights. With most companies movingtowards a data-driven culture, it's essential that they're able to analyse information from different sources.What if you could easily analyse all your product reviews from sites like Capterra or G2 Crowd? You'll beable to get real-time knowledge of what your users are saying and how they feel about your product.

The purpose of this project is todescribe some of the ways effective presentation of information from text, enabled byInformation Extraction, can improve analysis.

Benefits of Character Recognition:

1.Net Promoter Score (NPS) is one of the most popular customer satisfaction surveys. The first part of the survey asks the question: *"How likely are you to recommend [brand] to a friend?"* and needs to be answered with a score from 0 to 10. The results allow classifying customers into *promoters*, *passives*, and *detractors*.

2.Text mining can be very useful to analyse interactions with customers through different channels, like chat conversations, support tickets, emails.

3.Text mining can be useful to analyse all kinds of open-ended surveys such as post-purchase surveys or usability surveys.

4.This project canhelp companies become more productive, gain a better understanding of their customers, and use insights to make data-driven decisions.

5.Unlocking 'hidden' information and developing new knowledge

6. Improving research process and quality

PROJECT SOLUTION

Here we are going to describe the choices that were needed to be made before starting to

implement the solution.

There are 7 basic steps involved in preparing an unstructured text document for deeper analysis:

1. Language Identification.

2. Tokenization.

3. Sentence Breaking.

4. TFidf Transformation.

5. Modelling.

6. Visualizing errors.

7. Generating test predictions.

The first step to get up and running with text mining is gathering your data..

Data can be internal (interactions through chats, emails, surveys, spreadsheets, databases, etc) or external (information from social media, review sites, news outlets, and any other websites).

The second step is preparing your data. Text mining systems use several NLP techniques — like tokenization, stemming and stop removal — to build the inputs of your machine learning model.

The performance of a text classifier is measured through different parameters: accuracy, precision, recall and F1 score.

This section will go through the different metrics to analyse the performance of your text classifier, and explain how cross-validation works:

Accuracy indicates the number of correct predictions that the classifier has made divided by the total number of predictions.

Accuracy indicates the number of correct predictions that the classifier has made divided by the total number of predictions.

Precision evaluates the number of correct predictions made by the classifier, over the total number of predictions for a given tag (including both correct or incorrect predictions).

Recall indicates the number of texts that were predicted correctly, over the total number that should have been categorized with a given tag.

F1 score combines the parameters of precision and recall to give you an idea of how well your classifier is working.

Here we done the Text extraction from different methods called Regular Expression, Conditional Random Fields (CRF), and Evaluation.

We have named and described several Algorithms.

Naive Bayes (NB): They benefit from Bayes Theorem and probability theory to predict the tag of a text.

Support Vector Machines (SVM): This algorithm classifies vectors of tagged data into two different groups.

SCREENSHOTS

fig 1.1 represent the importing of packages thatcomes under libraries.

```
In [2]: #import required libraries

        import re
        import nltk

        import pandas as pd
        import numpy as np

        from bs4 import BeautifulSoup
        from nltk.corpus import stopwords
        from nltk.stem.porter import PorterStemmer
        english_stemmer=nltk.stem.SnowballStemmer('english')

        from sklearn.feature_selection.univariate_selection import SelectKBest, chi2, f_classif
        from sklearn.model_selection import train_test_split
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.linear_model import SGDClassifier, SGDRegressor
        from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
        from sklearn.naive_bayes import MultinomialNB
        from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
        import random
        import itertools

        import sys
        import os
        import argparse
        from sklearn.pipeline import Pipeline
        from scipy.sparse import csr_matrix
        from sklearn.feature_extraction.text import CountVectorizer
        import six
        from abc import ABCMeta
        from scipy import sparse
```

fig 1.1

fig 1.2 represents the importing of data

```
In [5]: #reading data from file
        data_file = 'C:\\Users\\HP\\Desktop\\Amazon_Unlocked_Mobile.csv'

        n = 413000  #total number of samples
        s = 20000   #desired sample size
        skip = sorted(random.sample(range(1,n),n-s)) #the 0-indexed header will not be included in the skip list


        data = pd.read_csv( data_file, delimiter = ",", skiprows = skip)
```

fig 1.2

Fig 1.3 represents the label exploration and showing much more than ratings.

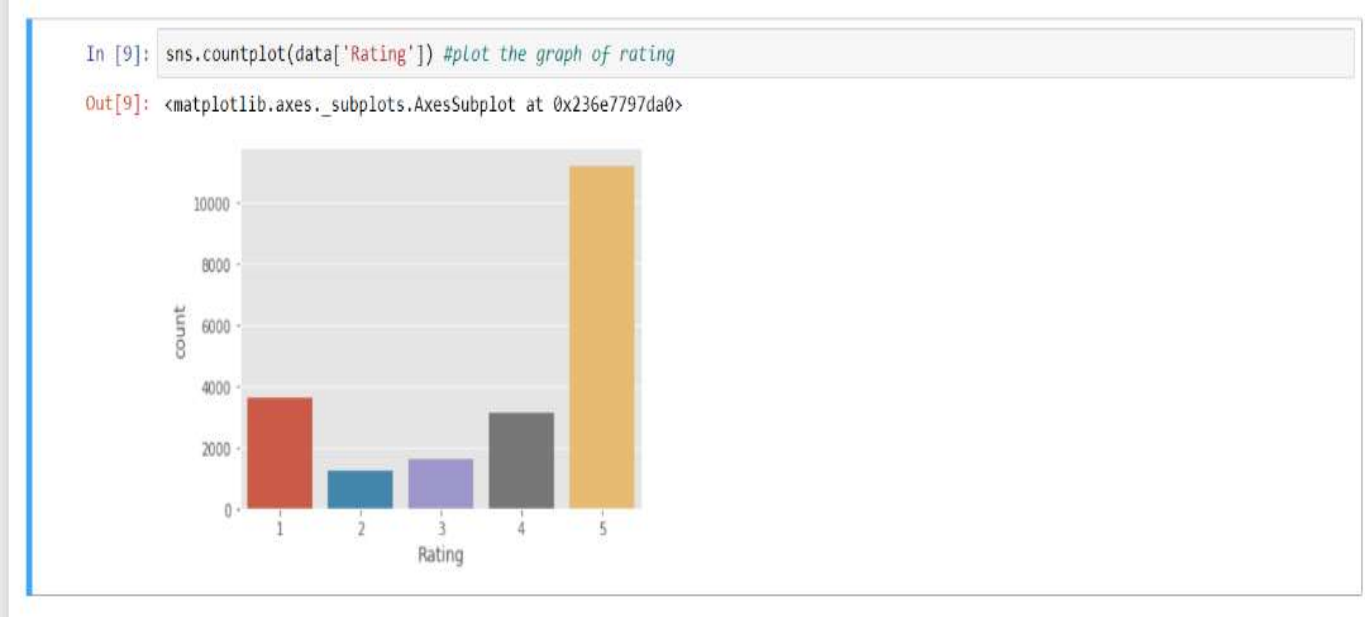


Fig 1.3

Fig 1.4 shows the TFidf transformation and with ngrams 1 and 4 also showing the selecting of best features for our prediction

```
In [11]: vectorizer = TfidfVectorizer( min_df=2, max_df=0.95, max_features = 200000, ngram_range = ( 1, 4 ),
                                       sublinear_tf = True )
         #Performs the TF-IDF transformation from a provided matrix of counts.

         vectorizer = vectorizer.fit(clean_train_reviews) #Learn vocabulary and idf from training set.
         train_features = vectorizer.transform(clean_train_reviews) #Transform documents to document-term matrix.

         test_features = vectorizer.transform(clean_test_reviews)  #Transform documents to document-term matrix.
```

```
In [12]: fselect = SelectKBest(chi2 , k=10000) #It removes all features whose variance doesn't meet some threshold
         train_features = fselect.fit_transform(train_features, train["Rating"]) #Fit to data, then transform it.
         test_features = fselect.transform(test_features) #Reduce test_features to the selected features.
```

Fig 1.4

Fig 1.5 shows the classification report of the project

```
In [17]: print(classification_report(test['Rating'], pred_2, target_names=['1','2','3','4','5']))
```

	precision	recall	f1-score	support
1	0.67	0.80	0.73	1035
2	0.50	0.13	0.21	356
3	0.49	0.15	0.22	501
4	0.42	0.18	0.25	970
5	0.73	0.95	0.83	3390
accuracy			0.69	6252
macro avg	0.56	0.44	0.45	6252
weighted avg	0.64	0.69	0.64	6252

Fig 1.5

Fig 1.6 represents the plotting of confusion matrix withou normalization.

```
In [20]: plot_confusion_matrix(cnf_matrix, classes=['1','2','3','4','5'],
                              title='Confusion matrix, without normalization')
```

```
Confusion matrix, without normalization
[[ 813   22   21   19  160]
 [ 162   51   19   31   93]
 [ 125   17   77   80  202]
 [  43    6   25  175  721]
 [  56    4   19  117 3194]]
```

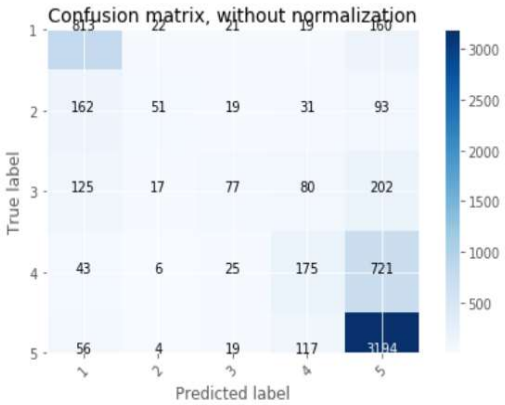


Fig 1.6

Fig 1.7 represents the generating of test predictions and also showing accuracy without epochs

```
Build model...

C:\Users\HP\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: UserWarning: The `dropout` argument is no longer support in `E
mbedding`. You can apply a `keras.layers.SpatialDropout1D` layer right after the `Embedding` layer to get the same behavior.
  This is separate from the ipykernel package so we can avoid doing imports until
C:\Users\HP\Anaconda3\lib\site-packages\ipykernel_launcher.py:4: UserWarning: Update your `LSTM` call to the Keras 2 API: `LSTM
(128, dropout=0.2, recurrent_dropout=0.2)`
  after removing the cwd from sys.path.

Train...

C:\Users\HP\Anaconda3\lib\site-packages\ipykernel_launcher.py:14: UserWarning: The `nb_epoch` argument in `fit` has been rename
d `epochs`.


WARNING:tensorflow:From C:\Users\HP\Anaconda3\lib\site-packages\tensorflow\python\ops\math_grad.py:1250: add_dispatch_support.<
locals>.wrapper (from tensorflow.python.ops.array_ops) is deprecated and will be removed in a future version.
Instructions for updating:
Use tf.where in 2.0, which has the same broadcast rule as np.where
Train on 14586 samples, validate on 6252 samples
Epoch 1/1
14586/14586 [==============================] - 111s 8ms/step - loss: 0.3376 - accuracy: 0.8643 - val_loss: 0.2985 - val_accurac
y: 0.8802
6252/6252 [==============================] - 11s 2ms/step
Test score: 0.2985385575415763
Test accuracy: 0.8802304267883301
Generating test predictions...
```

```
In [26]: print('prediction 7 accuracy: ', accuracy_score(test['Rating'], preds+1))

prediction 7 accuracy:  0.6666666666666666
```

Fig 1.7

RESULTS

Fig 1.8 shows that after adding epochs and proper testing of the model, it showing the more accurate with LSTM as per the input.

Then the output we get is the test accuracy which is 0.8873 i.e. 88%

```
Train on 14587 samples, validate on 6252 samples
Epoch 1/1
14587/14587 [==============================] - 34s 2ms/step - loss: 0.3296 - accuracy: 0.8685 - val_loss: 0.2814 - val_accurac
y: 0.8874
6252/6252 [==============================] - 1s 205us/step
Test score: 0.28144940024602894
Test accuracy: 0.8873957991600037
Generating test predictions...

In [30]: print('prediction 8 accuracy: ', accuracy_score(test['Rating'], preds+1))

prediction 8 accuracy:  0.6653870761356366
```

Fig 1.8

CONCLUSION

This project seeks patterns extraction from the analysis of large collections of documents in order to gain new knowledge. Its purpose is the discovery of interesting groups, trends, associations and the visualization of new findings.

Text mining is considering as a subset of data mining. For this reason, adopts text mining adopts the data mining techniques which uses machine learning algorithms. Computational linguistics techniques also provides techniques to text mining. This science studies natural language with computational methods to make them understandable by the operating system.

REFERENCES

[1] S. Cheriyan, S. Ibrahim, S. Mohanan and S. Treesa, "Intelligent Sales Prediction Using Machine Learning Techniques," 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), Southend, United Kingdom, 2018, pp. 53-58.

[2] M. Gurnani, Y. Korke, P. Shah, S. Udmale, V. Sambhe and S. Bhirud, "Forecasting of sales by using fusion of machine learning techniques".