GALGOTIAS UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SENTIMENT ANALYSIS USING MACHINE LEARNING

A Report for the Evaluation 3 of Project 2

Submitted by

Ambika

Admission No.:16SCSE101620

Enrollment No.:1613101112

in partial fulfillment for the award of the degree of

Bachelor Of Technology

IN

Computer Science &Engineering

School of Computing Science & Engineering

Under the Supervision of

Dr Jayakumar V, Associate Professor

APRIL / MAY- 2020

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **“SENTIMENT ANALYSIS USING MACHINE LEARNING”** is the bonafide work of “ **AMBIKA (1613101112)”** who carried out the project work under my supervision.

SIGNATURE	**SIGNATURE**
Dr. MUNISH SHABARWAL	**MR. S. JERALD NIRMAL KUMAR**
HEAD OFTHEDEPARTMENT	**SUPERVISOR**
PhD (Management), PhD (CS) **Professor & Dean,** **School of Computing Science & Engineering**	**Assoc. Prof.** **School of Computing Science & Engineering**

Abstract

Sentiment analysis or opinion mining is the computational study of people’s opinions, sentiments, attitude, and emotions expressed in written language. It is one of the most active research area in natural language processing and text mining in recent years. Its popularity is mainly due to two reasons. First, it has a wide range of applications because opinions are central to almost all human activities and are key influencer of our behaviours. Whenever we need to make a decision, we want to hear other’s opinion. Second, it present many challenging research problem, which has never been attempted before the year 2000. Part of the reason for the lack of study before was that there was little opinionated text in digital forms. It is thus no surprise that the inception and the rapid growth of the field coincide with those of the social media o the Web. In fact, the research has also spread outside of the computer science to management science and social sciences due to its importance to business and society as a whole. The approaches of the text sentiment analysis typically work at the particular level like phrases, sentences or document level. It aims at analyzing a solution for the sentiment classification at a fine-grained level, namely the sentence level in which polarity of sentence can be given by three categories as positive, negative and neutral.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	i
1	1. INTRODUCTION	1-3
	1.1 Purpose	3
	1.2 Literature Survey	4
	1.3 Design	5
	1.4 Output/Results/Screeshots	6-7
	1.5 Conclusion	8
	1.6 References	8

1. Introduction

Sentiment analysis refers to the use of natural processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topics or the overall contextual polarity of the document. The attitude may be of his or her judgment or evaluation effective state, or the intended emotional communication. Sentiment analysis is the process of detecting piece of writing for positive, negative or neutral feeling bound to it. Human have the innate ability to determine sentiment; however this process is time consuming , inconsistent and costly in business context. It's just not realistic to have people individually read tens of thousands of user customer reviews and score them for sentiments.

For example if we consider Semantria's cloud based sentiment analysis software. Semantria's cloud based sentiment analysis software extracts the sentiment of the document and its components through the following steps:

- A document is broken in its part of speech called POS tags, which identify the structural elements of a documents, paragraph, or sentences(i.e. Nouns, adjective, verbs, adverbs)
- Sentiment-bearing phrases, such as "terrible service", are identified through the use of specifically designed algorithm.
- Each sentiment-bearing phrase is a document is given a score based on a logarithms scale that ranges between -10 to 10.
- Finally, the score are combined to determine the overall sentiment of the document or the sentence document range between -2 to 2.

Semantria's cloud based sentiment analysis software is based upon the Natural Language Processing and delivers more consistent results than two humans. Using automated sentiment analysis, Semantria analyzes each document and its components based on sophisticated algorithms developed to extract sentiments from your content in a similar manner as a human –only 60,000 times faster.

Existing approaches to sentiment analysis can be grouped into three main categories:

- Keyword spotting
- Lexical affinity
- Statistical methods

Keyword spotting is the most naïve approach and probably also most popular because of its accessibility and economy. Text is classified into affect categories based on the presence of fairly unambiguous affect words like 'happy', 'sad', 'afraid', and 'bored'. The weakness of this approach in two areas: poor recognition of affect when negation is involved and reliance on surface features. About its first weakness, while the approach can correctly classify the sentence "today was a happy day" as being happy, it is likely to fail on a sentence like "today wasn't a happy day". About its second weakness, the approach relies on the presence of obvious affect words that are only surface features of the prose.

In practice, lot of sentences convey affect through underlying meaning rather than affect adjective. For example, the text "My husband just filed for divorce and wants to take custody of my children away from me" certainly evokes strong emotions, but use not affect keywords, and therefore, cannot be classified using keyword spotting approach.

Lexical affinity is slightly more sophisticated than keyword spotting as , rather than simply detecting obvious affect words, it assigns arbitrary words a probabilistic 'affinity' for a particular emotion. For example, ‘accident’ might assign 75% probability of being identified as negative affect, as in ‘car accident’ or ‘hurt by accident’. These probabilities are usually trained from linguistic corpora. Though often outperforming pure keyword spotting, there are two main problem with this approach first lexical affinity, operates solely on the word-level, can easily be tricked by sentences like “I avoid an accident” (negation) and “I met a girl by accident” (other word sense) Second, lexical affinity probabilities are often biased towards text of a particular genre, dictated by the source of the linguistic corpora. This make it difficult to develop a reusable, domain-independent model.

Statistical methods, such as Bayesian inference and support vector machine, have been popular for the affect classification of texts. By feeding a machine learning algorithm a large training corpus of affectively annotated texts, it is possible for the system to not only learn the affective valence of affect words, but also to take into account the valence of arbitrary words, punctuations, and words co-occurrence frequencies. However, traditional statistical methods are generally semantically weak, meaning that, with exception of obvious affect keywords, other lexical or co-occurrence elements in a statistical model have little predictive value individually. As a result, statistical text classifiers only works with the acceptable accuracy when given a sufficiently large text input. So, while these methods may be able to affectively classify user’s text on the page or paragraph level, they do not work on smaller texts units such as sentence or clauses.

1.1 PURPOSE

Sentiment classification is a way to analyze the subjective information in the text and then mine the opinion. Sentiment analysis is a procedure by which information is extracted from the opinions, appraisals and emotions of the people in regards to entities, events and their attributes.

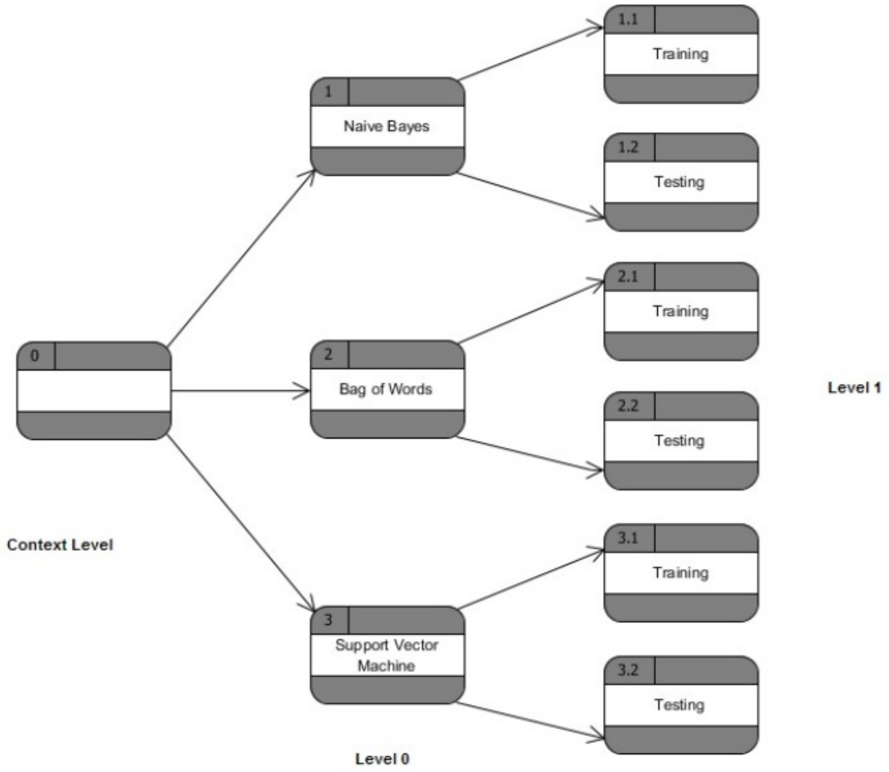
1.2 Literature Survey

The most fundamental problem in sentiment analysis is the sentiment polarity categorization, by considering a dataset containing over 5.1 million product reviews from Amazon.com with the products belonging to four categories. A max-entropy POS tagger is used in order to classify the words of the sentence, an additional python program to speed up the process. The negation words like no, not, and more are included in the adverbs whereas Negation of Adjective and Negation of Verb are specially used to identify the phrases. The following are the various classification models which are selected for categorization: Naïve Bayesian, Random Forest, Logistic Regression and Support Vector Machine. For feature selection, Pang and Lee suggested to remove objective sentences by extracting subjective ones. They proposed a text-categorization technique that is able to identify subjective content using minimum cut. Gann et al. selected 6,799 tokens based on Twitter data, where each token is assigned a sentiment score, namely TSI (Total Sentiment Index), featuring itself as a positive token or a negative token. Specifically, a TSI for a certain token is computed as:

$$TSI = \frac{p - \frac{tp}{tn} \times n}{p + \frac{tp}{tn} * n}$$

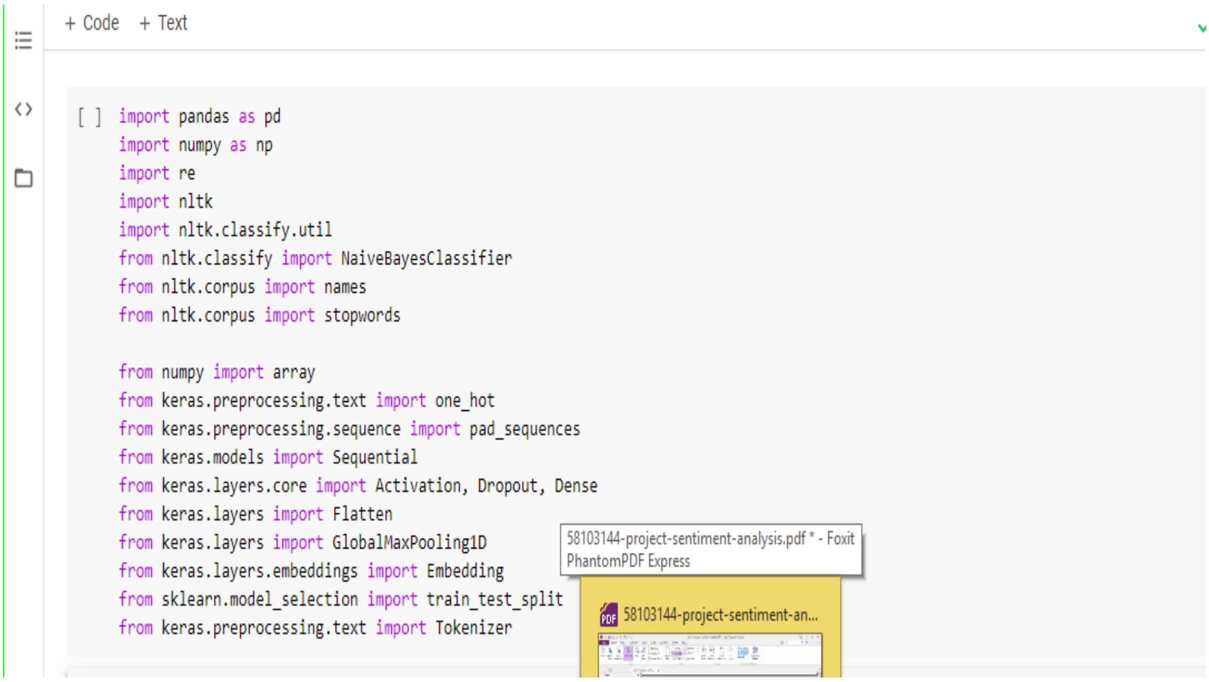
where p is the number of times a token appears in positive tweets and n is the number of times a token appears in negative tweets is the ratio of total number of positive tweets over total number of negative tweets.

1.3 UML DIAGRAM



1.4 OUTPUTS/RESULTS/SCREENSHOTS

Importing Required Libraries

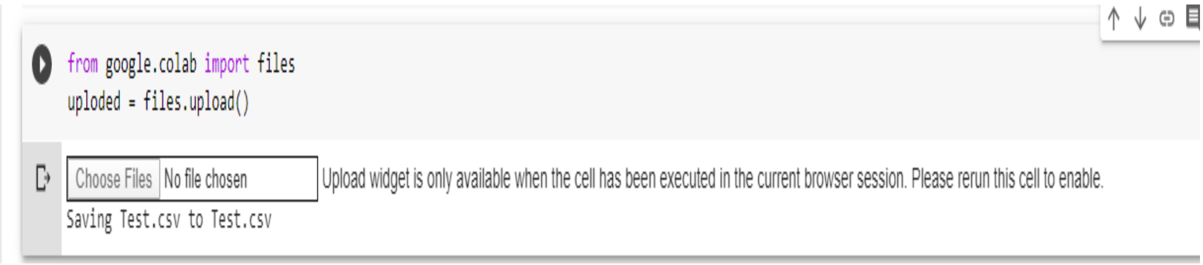

+ Code
+ Text
[ ]
58103144-project-sentiment-analysis.pdf * - Foxit PhantomPDF Express
58103144-project-sentiment-an...

```
import pandas as pd
import numpy as np
import re
import nltk
import nltk.classify.util
from nltk.classify import NaiveBayesClassifier
from nltk.corpus import names
from nltk.corpus import stopwords

from numpy import array
from keras.preprocessing.text import one_hot
from keras.preprocessing.sequence import pad_sequences
from keras.models import Sequential
from keras.layers.core import Activation, Dropout, Dense
from keras.layers import Flatten
from keras.layers import GlobalMaxPooling1D
from keras.layers.embeddings import Embedding
from sklearn.model_selection import train_test_split
from keras.preprocessing.text import Tokenizer
```

Importing and Analyzing the Dataset

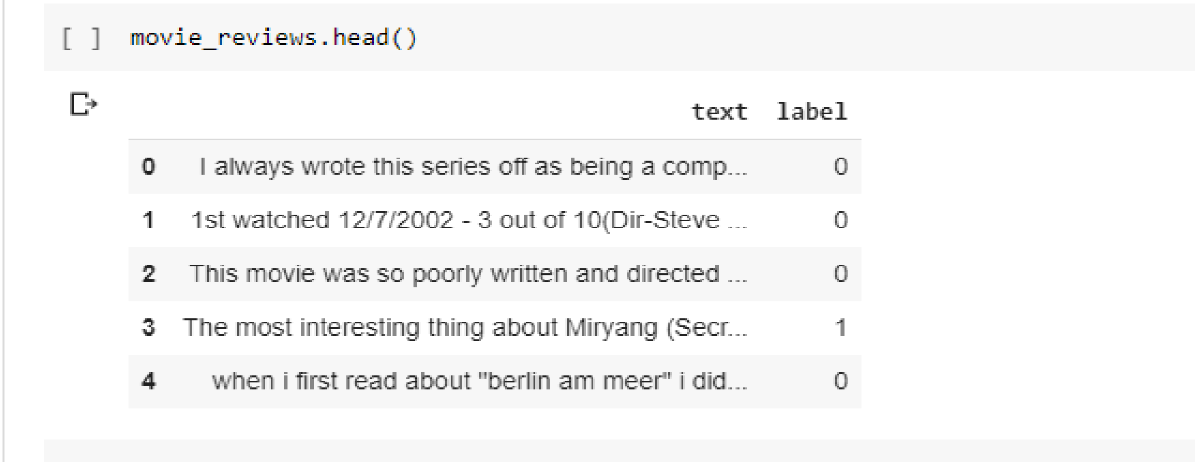
In the script read_csv() method of the pandas library to read the CSV file containing our dataset.



```
from google.colab import files
uploded = files.upload()
```

```
Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving Test.csv to Test.csv
```

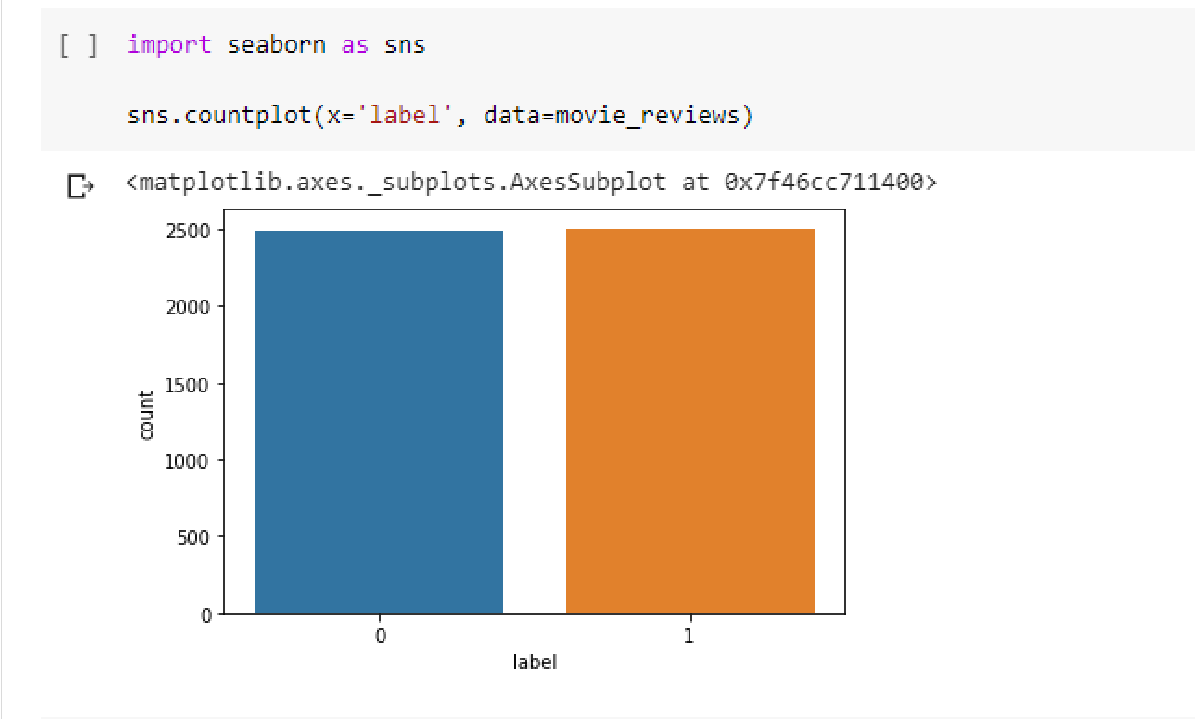
Print the first 5 rows of the dataset using the head() method.



```
[ ] movie_reviews.head()
```

	text	label
0	I always wrote this series off as being a comp...	0
1	1st watched 12/7/2002 - 3 out of 10(Dir-Steve ...	0
2	This movie was so poorly written and directed ...	0
3	The most interesting thing about Miryang (Secr...	1
4	when i first read about "berlin am meer" i did...	0

The distribution of positive and negative sentiments in dataset.



```
[ ] import seaborn as sns

    sns.countplot(x='label', data=movie_reviews)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f46cc711400>
```

1.5 CONCLUSION

Sentiment analysis is an emerging research area in text mining and computational linguistics, and has attracted considerable research attention in the past few years. Future research shall explore sophisticated methods for opinion and product feature extraction, as well as new classification models that can address the ordered labels property in rating inference. Applications that utilize results from sentiment analysis is also expected to emerge in the near future.

1.6 REFERENCES

- https://www.stackabuse.com/python-for-nlp
- https://www.colab.research.google.com