JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Determining the polarity and statistics of chatbased on sentiment analysis

Submitted By
Aditya kumar
Mahadev kumar
Deeshu sharawat

Under The Supervision ofDr.Pooja singh
Asst. prof

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
INDIA
JUNE,2022-23

1 :INTRODUCTION

1.1 : Problem Definition

Automate detection of different sentiments from textual comments and feedback, A machine learning model is created to understand the sentiments of the restaurant reviews. The problem is that the review is in a textual form and the model should understand the sentiment of the review and automate a result.

The main motive behind this project is to classify whether the given feedback or review in textual context is positive or negative. Reviews can be given to the model and it classifies the review as a negative review or a positive. This shows the satisfaction of the customer or the experience the customer has experienced.

The basic approach was trying a different machine learning model and look for the one who is performing better on that data set. The restaurant reviews are very related to the project topic as reviews are made on websites and we can apply this model on such data sets to get the sentiments.

1.2 :Project Overview/Specifications*1.2.1: Task

Definition:

To develop a machine learning model to detect different types of sentiments contained in a collection of English sentences or a large paragraph. I have chosen Restaurant reviews as my topic. Thus, the objective of the model is to correctly identify the sentiments of the users by reviews which is an Englishparagraph and the result will be in positive or negative only.
For example,
If the review given by the user is:
" We had lunch here a few times while on the island visiting family and friends. The servers here are just wonderful and have great memories it seems. We sat on the oceanfront patio and enjoyed the view with our delicious wine and lunch. Must try! "
Then the model should detect that this is a positive review. Thus the output for this text will be Positive.

1.2.2 : Algorithm Definition:

The data set which I chose for this problem is available on Kaggle. The sentiment analysis is a classification because the output should be either positive or negative. That is why I tried 3 of the classification algorithms on this data set.
Multinomial Naive BayesBernoulli Naive Bayes

Logistic Regression:

i) Multinomial Naive Bayes: Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature. Bayes theorem calculates probability P(c|x) where c is the class of the possible outcomes and x is the given instance which has to be classified, representing some certain features.
$P(c|x) = P(x|c) * P(c) / P(x)$
Naive Bayes is mostly used in natural language processing (NLP) problems. Naive Bayes predict the tag of a text. They calculate the probability of each tag for a given text and then output the tag with the highest one.

ii) Bernoulli Naive Bayes: BernoulliNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input (depending on the binarize parameter). The decision rule for Bernoulli naive Bayes is based on:- $P(x_i \mid y) = P(i \mid y)\, x_i + (1 - P(i \mid y))\,(1 - x_i)$ which differs from multinomial NB's rule in that it explicitly penalizes the non- occurrence of a feature that is an indicator for class, where the multinomial variant would simply ignore a non-occurring feature.

In the case of text classification, word occurrence vectors (rather than word count vectors) may be used to train and use this classifier. BernoulliNB might perform better on some datasets, especially those with shorter documents. It isadvisable to evaluate both models if time permits.

iii) Logistic Regression Logistic regression is a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for the given set of features (or inputs),X.

Contrary to popular belief, logistic regression is a regression model. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as "1". Just like Linear regression assumes that the data follows a linear function, Logistic regression models the data usingthe sigmoid function.

LIBRARIES USED:

NLTK:

NLTK (Natural Language Toolkit) Library is a suite that contains libraries and programs for statistical language processing. It is one of the most powerful NLP libraries, which contains packages to make machines understand human language and reply to it with an appropriate response.

PANDAS :

Pandas is an open-source, BSD-licensed Python library providing highperformance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

SKLEARN:

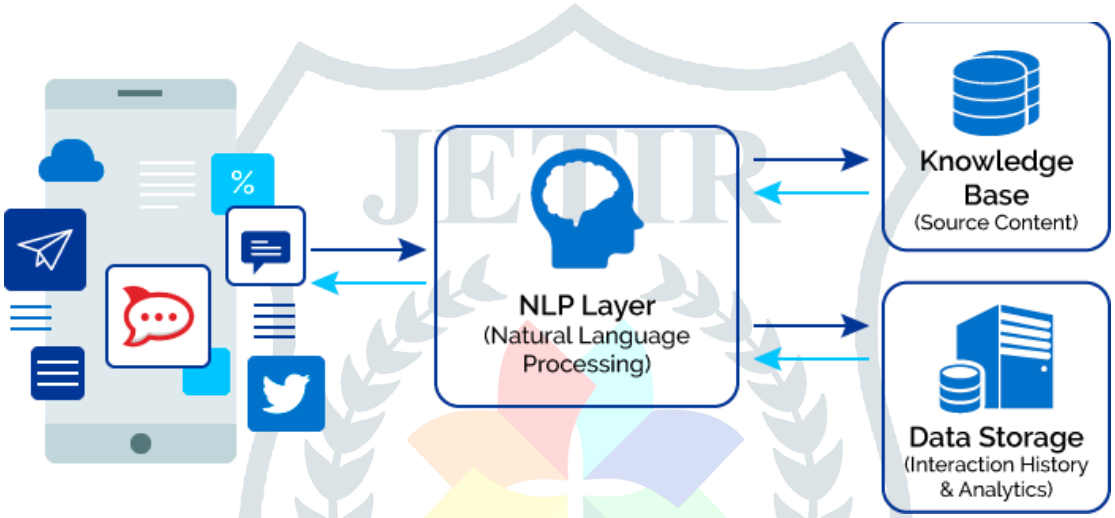
Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

NUMPY:

NumPy is a Python library used for working with arrays.
It also has functions for working in domain of linear algebra, Fourier transform,and matrices.
NumPy was created in 2005 by Travis Oliphant. It is an open-source project andyou can use it freely.
NumPy stands for Numerical Python.

FLASK:

NumPy is a Python library used for working with arrays.
It also has functions for working in domain of linear algebra, Fourier transform,and matrices.
NumPy was created in 2005 by Travis Oliphant. It is an open-source project andyou can use it freely.
NumPy stands for Numerical Python.



1.3 Hardware Specifications

PC with 8 GB RAM And Sufficient storage.

1.4 Software Specifications

Python, Flask, PyCharm IDE

2: LITERATURE REVIEW:

Some of the active researches on NLP phenomena include the Syntactic phenomena: those that pertain to the structure of a sentence and the order of words in the sentence, based on the grammatical classes of words rather than their meaning (e.g. discriminative models for scoring parses, coarse to fine efficient approximate parsing, dependency grammar); Machine translation (e.g. models and algorithms, low- resource and morphological complex language); Semantic phenomena : those that pertain to the meeting of a sentence relatively independent of the context in which the language occurs(e.g. sentiment analysis, summarization, information extraction ,slot filling, discourse analysis, textual entailment);Pragmatic phenomena such as Speech: those that relate the meaning of a sentence to the context in which it occurs. This context can be linguistic (such as the previous text or dialogue) or, non-linguistic (such as knowledge about person who produced the language, about goals of the communication, about the objects in the current visual field, etc. (e.g. language modelling-syntax and semantics, models of acoustics, pronunciation) [17]; [18]. Speech recognition and information retrievalhave finally gone commercial and there is a ton of text and speech on the Internet, cell phones, etc. It is now clear that studies regarding anything about a language are possible, e.g. formalizing some insights

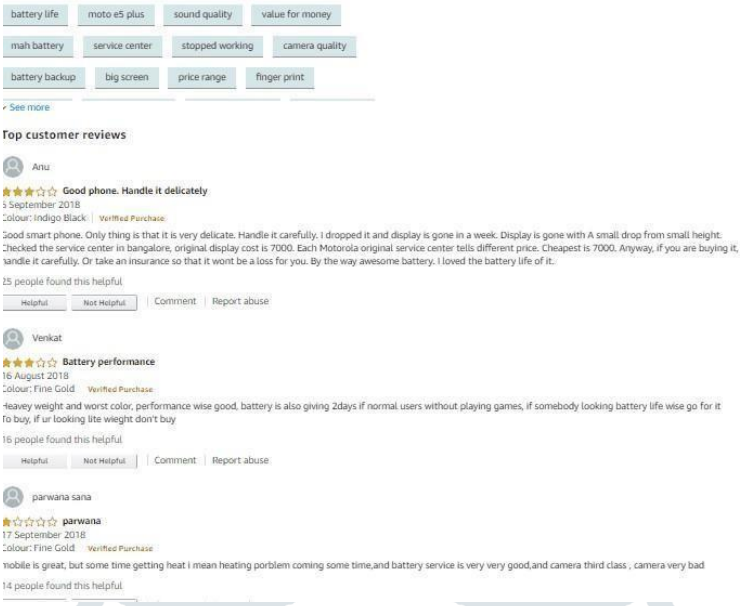
e.g. discrete knowledge (what is possible) and continuous knowledge (what is likey); studying the formalism mathematically; developing and implementing algorithms and testing on real data. The current and on- going future changes or improvements which need to be done to NLP are: to add features to existing interfaces, back end processing should befully implemented (e.g. information extraction and normalization to build databases. Another anticipated improvement is of having hand held devices with translators and personal conversation recorder with topical searches [17].

3: PROBLEM FORMULATION:

Online product reviews are a great source of information for consumers. From the sellers' point of view, online reviews can be usedto gauge the consumers' feedback on the products or services they areselling.

However, since these online reviews are quite often overwhelming in terms of numbers and information, an intelligent system, capable of finding key insights (topics) from these reviews, will be of great help for both the consumers and the sellers. This system will serve two purposes:

1. Enable consumers to quickly extract the key topics covered bythereviews without having to go through all of them
2. Help the sellers/retailers get consumer feedback in the formoftopics (extracted from the consumer reviews)


battery life | moto e5 plus | sound quality | value for money

mah battery | service center | stopped working | camera quality

battery backup | big screen | price range | finger print

See more

Top customer reviews

Anu

Good phone. Handle it delicately

5 September 2018

Colour: Indigo Black | Verified Purchase

Good smart phone. Only thing is that it is very delicate. Handle it carefully. I dropped it and display is gone in a week. Display is gone with A small drop from small height. Checked the service center in bangalore, original display cost is 7000. Each Motorola original service center tells different price. Cheapest is 7000. Anyway, if you are buying it, handle it carefully. Or take an insurance so that it wont be a loss for you. By the way awesome battery. I loved the battery life of it.

25 people found this helpful

Helpful | Not Helpful | Comment | Report abuse

Venkat

Battery performance

16 August 2018

Colour: Fine Gold Verified Purchase

Heavey weight and worst color, performance wise good, battery is also giving 2days if normal users without playing games, if somebody looking battery life wise go for it fo buy, if ur looking lite wieght don't buy

16 people found this helpful

Helpful | Not Helpful | Comment | Report abuse

parwana sana

parwana

17 September 2018

Colour: Fine Gold Verified Purchase

mobile is great, but some time getting heat i mean heating porblem coming some time,and battery service is very very good,and camera third class , camera very bad

14 people found this helpful

1.5 Literature Review Summary

Table 2.1: Literature review summary

Year and citation	Article Title	Purpose of the study	Tools/ Software used	Comparison of technique done	Source (Journal/ Conference)	Findings	Data set (if used)	Evaluation parameters
2010								

4: OBJECTIVES

Based on document analysis, this paper summarizes the information on NLP, the general overview, history, and previous works on NLP. It then considers applications of NLP. The challenges and failures of NLP together with current and future research of NLP are also discussed briefly in this paper. The research paper is intended to give an understating to researchers, scholarly peers and companies who wish to stay abreast with the NLP technologies and applications from the past, present and future.

5: METHODOLOGY:

All the models were judged based on a few criteria. These criteria are also recommended by the scikit-learn website itself for the classification algorithms.The criteria are:

Accuracy score: Classification Accuracy is what we usually mean when we use the term accuracy. It is the ratio of the number of correct predictions to the totalnumber of input samples.

Confusion Matrix: A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. i) There are two possible predicted classes: "yes" and "no". If we were predicting the presence of a disease, for example, "yes" would mean they have the disease, and "no" would mean they don't have the disease. ii) The classifier made a total of 165 predictions (e.g., 165 patients were being tested for the presence of that disease). iii) Out of those 165 cases, the classifier predicted "yes" 110 times, and "no" 55 times. iv) In reality, 105 patients in the sample have the disease, and 60 patients do not.
true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease. true negatives (TN): We predicted no,and they don't have the disease.
false positives (FP): We predicted yes, but they don't have the disease. (Alsoknown as a "Type I error.")
false negatives (FN): We predicted no, but they do have the disease. (Also knownas a "Type II error.")

F1 score F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but lower recall, gives you an extremely

accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as : F1 Score tries to find the balance between precision and recall.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier.

Recall: It is the number of correct positive results divided by the number of allrelevant samples (all samples that should have been identified as positive).

6: RESULTS AND DISCUSSION

The motive of the model is to correctly detect the sentiments of the textual reviews or feedback. The developed model has an accuracy of 77.67% and successfully detects the sentiments of the textual reviews or feedback. The model has been tested with few of the online reviews and was found that it detects the sentiments correctly. Thus, can conclude that the motive was successful and the model can be used to detect the sentiments of the reviews and feedback.
All of the 3 mentioned machine learning models are very measured on the abovementioned metrics. The result of the evaluation of the metrics is mentioned below:

i) Multinomial Naive Bayes:
Confusion Matrix: [[119, 33], [34, 114]]
Accuracy, Precision and Recall Accuracy is 77.67 % Precision is 0.78 Recall is0.77

ii) Bernoulli Naive Bayes
Confusion Matrix: [[115, 37], [32, 116]]
Accuracy, Precision and Recall Accuracy is 77.0 % Precision is 0.76 Recall is 0.78

iii) Logistic Regression

Confusion matrix: [[125, 27], [43, 105]]
Accuracy, Precision and Recall Accuracy is 76.67 % Precision is 0.8 Recall is 0.71 The above results are very clear. That is why I chose Multinomial Naive Bayes and tried to tweak it by tuning the model for better results. For which I have iterated with different parameters and found the best-suited parameter with the highest accuracy. Let's now define the most basic terms, which are whole numbers (not rates):

- true positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
- true negatives (TN): We predicted no, and they don't have thedisease.
- false positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")
- false negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

CHAPTER 7: CONCLUSION AND FUTURE SCOPE

There is always a scope of improvement. Here are a few things which can be considered to improve.

Different classifier models can also be tested.
The remaining two models can be tuned for better results.
For example, after plotting. AUC in Logistic Regression we may getbetter results.

Try a different data set. Sometimes a data set plays a crucial role too. Some other tuning parameters to improve the accuracy of the model.

REFERENCES:

Bibliography

https://www.geeksforgeeks.org/applying-multinomial-naive-bayes-to- nlpproblems/

https://www.geeksforgeeks.org/understanding-logistic-regression/

https://www.geeksforgeeks.org/naive-bayes-classifiers/

https://towardsdatascience.com/metrics-to-evaluate-your-machine- learningalgorithm-f10ba6e38234

https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

https://www.analyticsvidhya.com/blog/2017/09/common-machine- learningalgorithms/

https://www.kingandprince.com/dining-guest-reviews.aspx

https://www.tripadvisor.in/Restaurant_Review-g1078423-d948529- ReviewsMartin_Berasategui-Lasarte_Province_of_Guipuzcoa_Basque_Country.html