**UML501 Machine Learning**

**CLASSIFY REVIEW POSITIVE OR NEGATIVE**

**USING NLP AND FLASK**

**Submitted by:**

**Name – Manavsharma**

**Roll No – 101703320**

**Group – COE 15**

**Submitted to:**

**Dr. Parteek Bhatia**



**Computer Science and Engineering Department**

**TIET, Patiala**

**November 2019**

**PROBLEM STATEMENT**

To classify whether the given review about the restaurant in the feedback is positive(good) or negative(not good).

Input taken is the text which is the review and then output on the screen is the result which is positive or negative.

**DATASET**

I picked the dataset for the problem from superdatascience.

Link to the dataset**:**

[**https://www.superdatascience.com/pages/machine-learning**](https://www.superdatascience.com/pages/machine-learning)

**DESCRIPTION OF FIELDS**

It contains 1000 rows and two columns one is of review,and the second is 1 or 0.

1 means positive review and 0 means negative review.

**DATA PREPROCESSING**

Tokenization: This is a process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens. The list of tokens becomes input for further processing. NLTK Library has word\_tokenize and sent\_tokenize to easily break a stream of text into a list of words or sentences, respectively.

Word Stemming/Lemmatization: The aim of both processes is the same, reducing the inflectional forms of each word into a common base or root. Lemmatization is closely related to stemming. The difference is that a stemmer operates on a single word without knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. However, stemmers are typically easier to implement and run faster, and the reduced accuracy may not matter for some applications.

**MACHINE LEARNING MODELS APPLIED**

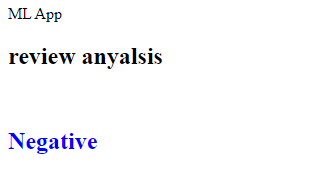
.

Bag-of-words model(BoW ): is the simplest way of extracting features from the text. It converts text into the matrix of occurrence of words within a document. This model concerns about whether given words occurred or not in the document.

We generate document term matrix by using scikit-learn's= CountVectorizer.

Naïve Bayes:Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

**RESULTS**



**CONCLUSION**

**From this model trained from the dataset is giving accuracy of 73% and can be improved by taking large dataset.**