CHAPTER 1

INTRODUCTION

Twitter is a microblogging website where people can share their feelings by sending tweets. Using twitter we can get overall opinion on anything. The amount of data accumulated on twitter is very huge. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm.

Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data.

The aim of this project is used to design web application which is used to analyze tweets. Here we will be performing sentimental analysis on tweets and produce results either positive or negative. The main feature of this web application is that it helps to determine the opinion about people’s product, or politician or any other by analyzing the tweets.

* 1. **PYTHON**

Python is a high level, dynamic programming language which is used for this project. It is interpreted language which makes the testing and debugging quickly as there is no compilation step. Python is simple, very powerful and dynamic programming language.

* 1. **NLTK**

Natural Language Toolkit (NLTK) is a library in python, which provides a base for building programs and classification of data. NLTK is a collection of resources for Python that can be used for text processing, classification, tagging and tokenization. NLTK provides various functions which are used in pre-processing of data so that data available from twitter become fit for mining and extracting features.

In our thesis we use Python as our base programming language which is used for writing code snippets. NLTK is a library of Python which plays a very important role in converting natural language text to a sentiment either positive or negative.

* 1. **MACHINE LEARNING**

Machine learning is a field in computer science aiming to imitate the human learning process. Machine Learning is a branch of Artificial Intelligence where computer learns from the data (past experiences) and makes future prediction. It finds the pattern in data, based on pattern it predicts for unseen data.

Machine Learning classified into 3 types:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning

Supervised Machine Learning

Supervised Learning algorithms learn from labeled input data, where labeled input data associated with output responses that consist of numeric values or string labels. Based on Output Variable Supervised Machine Learning classified into two types:

1. Regression:

If output variable is continuous then these types of problems can be considered as regression type.

Example: Predicting salary of person based on past experience of person.

1. Classification:

If output variable is categorical then these types of problems can be considered as classification type.

This problem is classification problem. Here we will use Logistic Regression, Random Forest, Gradient Boosting, XG Boost, KNN algorithms and we will compare and will take algorithm which will give best.

Example: Predicting whether patient is having cancer or not.

Unsupervised Machine Learning

If Machine Learning learns from unlabeled data, where data is not associated with target response.

Algorithms:

1. Clustering Algorithms
2. Association Clustering

Reinforcement Machine Learning

Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task.

CHAPTER 2

PROBLEM STATEMENT

Twitter is a microblogging website where people can share their feelings by sending tweets. Using twitter we can get overall opinion on anything. The amount of data accumulated on twitter is very huge. Twitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negative or neutral with the help of different machine learning algorithm.

Sentimental Analysis is done by using various machine learning techniques, statistical models and Natural Language Processing (NLP) for the extraction of feature from a large data.

The aim of this project is used to design web application which is used to analyze tweets. Here we will be performing sentimental analysis on tweets and produce results either positive or negative. The main feature of this web application is that it helps to determine the opinion about people’s product, or politician or any other by analyzing the tweets.

CHAPTER 3

DATASET

The dataset available on Analytics Vidhya. Dataset contains two columns one is independent feature that is tweet and one is dependent feature is label.

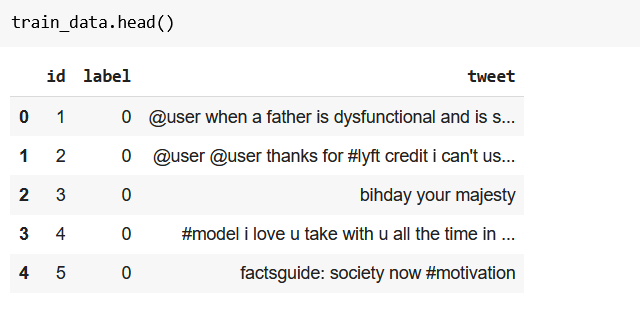


FIG 3.1: Dataset

This dataset has 31962 observations and two columns. Here we need to predict label columns. In this label is numerical type and tweet is categorical type which is in the form of strings.

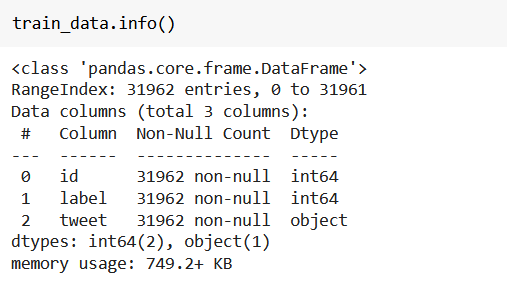


FIG 3.2: Dataset Information

CHAPTER 4

SOFTWARE REQUIREMENTS SPECIFICATION

4.1 INTRODUCTION:

Software Requirement Specification is the beginning point of the software developing activity. As system grew more complex it became evident that the goal of the entire system cannot be easily understood. Hence the need for the requirement phase arises. The Software Requirements Specification is one of the means of translating the ideas of the minds of clients (the input) into format document (the output of the requirement phase).

4.2 PURPOSE:

The aim of this project is used to design web application which is used to analyze tweets. Here we will be performing sentimental analysis on tweets and produce results either positive or negative.

4.3 REQUIREMENTS:

4.3.1 HARDWARE REQUIREMENTS

• Processor - Pentium –III

• Speed - 1.1 GHz

• RAM - 256 MB (min)

• Hard Disk - 260 GB

4.3.2 SOFTWARE REQUIREMENTS

* Operating System – Windows10/98/2000/XP
* Web Framework : Flask
* Programming Language: Python.
* Libraries: Sklearn, NLTK

CHAPTER 5

SYSTEM IMPLEMENTATION

5.1 Modules

* + 1. Data Preprocessing

5.1.2 Model Training

5.1.1 Data Preprocessing:

There are several steps in Data preprocessing:

1. Removing unwanted characters.
2. Removing Stop words.
3. Lemmatization.
4. Treating imbalanced dataset.
5. Converting Text into numerical forms.

Removing unwanted characters:

Lower() is used to make string to lower case. And regular expression is used to remove unwanted characters.

Removing stop words:

Stop words does not contain much information compared to other words. We will use NLTK library to remove stop words.

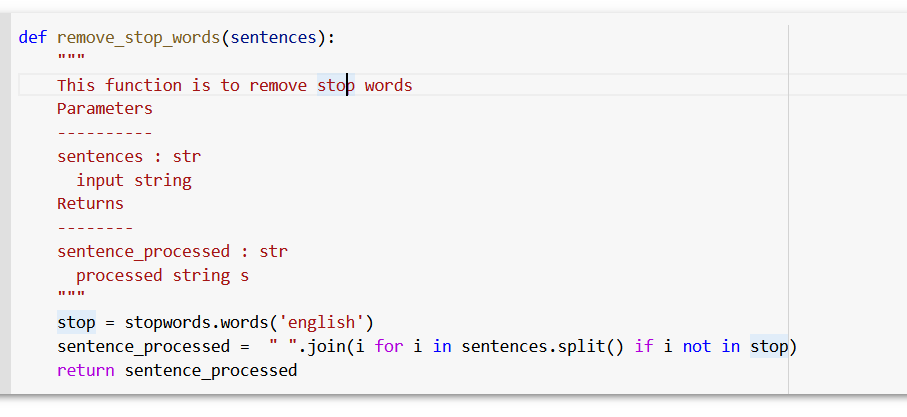


FIG 5.1: Removing Stop Words

Lemmatization:

Lemmatization is used to get root words.

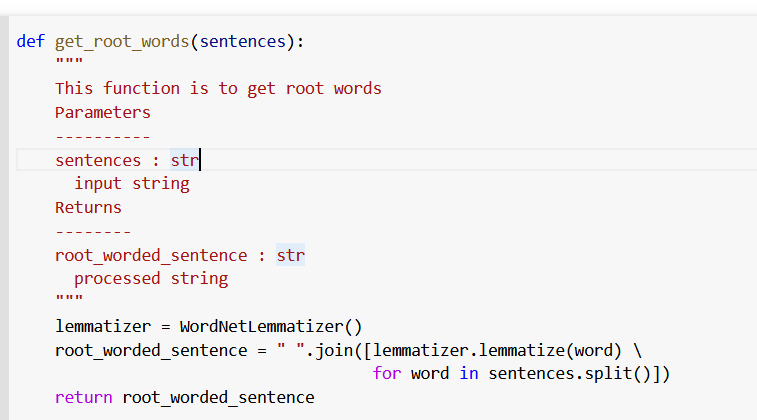


FIG 5.2: Lemmatizing

Imbalanced Dataset Treatment:

This dataset is imbalanced; we will use Sklearn to treat imbalanced dataset.

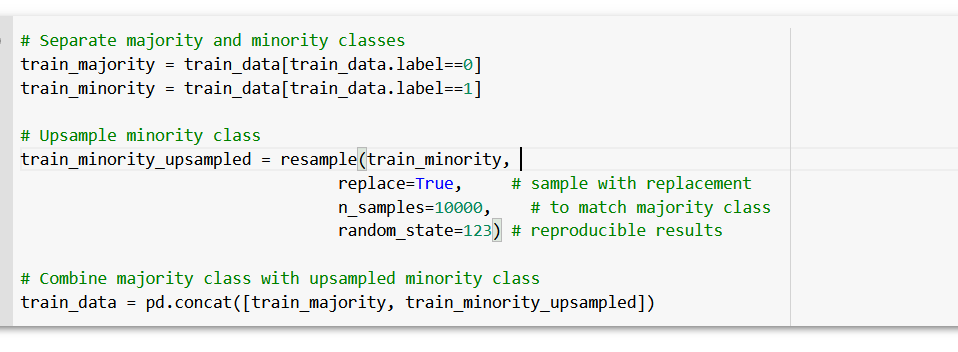


FIG 5.3: Treating Imbalanced Dataset

Converting Text into numerical forms:

Converting textual data into numerical data. Here we will use TF-IDF Vectorizer to convert text into numerical.



FIG 5.4: Vectorizer

5.1.2 Model Training

We will train various Machine Learning algorithms and we will compare all with the results and we will select the best one.

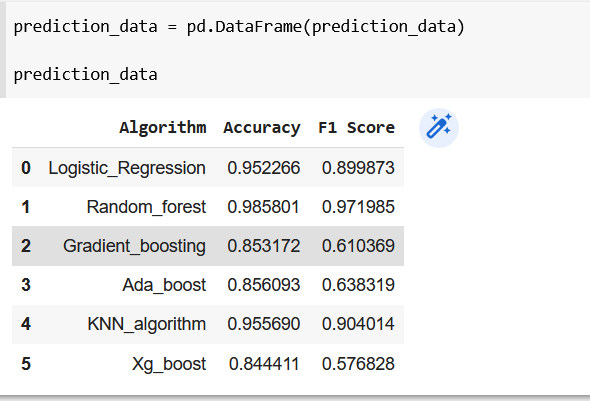


FIG 5.5: Algorithms withAccuracies

CHAPTER 6

TESTING

6.1 SOFTWARE TESTING

GENERAL

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, subassemblies, assemblies and/or a finished product it is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

DEVELOPING METHODOLOGIES

The test process is initiated by developing a comprehensive plan to test the general functionality and special features on a variety of platform combinations. Strict quality control procedures are used. The process verifies that the application meets the requirements specified in the system requirements document and is bug free. The following are the considerations used to develop the framework from developing the testing methodologies.

6.2 Types of Tests

6.2.1 Unit testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program input produces valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This Is structural testing, that relies on knowledge of its construction and is invasive configuration.

6.2.2 Functional test

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input: identified classes of valid input must be accepted.

Invalid Input: identified classes of invalid input must be rejected.

Functions: identified functions must be exercised.

Output: identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

6.2.3 System Test

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

6.2.4 Performance Test

The Performance test ensures that the output is produced within the time limits, and the time taken by the system for compiling, giving response to the users and request being send to the system for to retrieve the results.

6.2.5 Integration Testing

Software integration testing is the incremental integration testing of two or more integrated software components on a single platform to produce failures caused by interface defects. The task of the integration test is to check that components or software applications.

CHAPTER 7

SNAPSHOT

