ANNA UNIVERSITY, CHENNAI – 600 025.

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

MASTER OF COMPUTER APPLICATION (SS) 3 – YEAR PROGRAM

BATCH – 2

ZEROth REVIEW REPORT

ON

SENTIMENT ANALYSIS ON SOCIAL MEDIA TO DETECT DEPRESSION USING MACHINE LEARNING AND DEEP LEARNING

GUIDED BY MS.R.L.JASMINE

SUBMITTED BY

PRIYANKA G

2019272030

ABSTRACT:

In today's society, the use of social media has become a necessary daily activity. It is a valuable communication tool with others locally and worldwide, as well as to share, creates, and spread information. Although social media certainly has several remarkable features, the demerits are undeniable as well. Recent studies have indicated a correlation between high usage of social media sites and increased depression. This project aims to detect a probable depressed Twitter user based on his/her network behavior and tweets. Machine Learning, Deep Learning and Natural Language Techniques are used for classification process and predicting whether the user is depressed or not.

INTRODUCTION:

Natural Language Processing

Natural Language Processing is a method that communicates with an intelligent system using a natural example, say English. It can be used to perform many tasks on these intelligent systems. It gives computers the ability to understand text and spoken words in much the same way human beings can. The exact meaning or the dictionary meaning of the text is extracted using the Semantic Analysis.

Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is an approach to natural language processing (NLP) that identifies the emotional tone behind a body of text. This is a popular way for organizations to determine and categorize opinions about a product, service, or idea. It depicts not only on polarity (positive, negative neutral) but also on emotions (happy, sad, angry, etc.). It uses various Natural Language Processing algorithms. It is the contextual mining of words that indicates the social sentiment of a brand.

Feature Extraction

Feature extraction reduces to the processing groups from the initial raw data. Feature extraction is the method of selecting and combining data into features reducing the data amount that must be accurately processed and the original data set described thoroughly. The amount of redundant data for a given analysis will also get reduced. The machine’s efforts invariable combinations (features)building and data reduction facilitate machine learning process by learning and generalization steps.

Machine Learning

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. It is the process of teaching a system on making accurate predictions while feeding data is. It shows the working of an algorithm which learns more accurate in its predictions.

Deep Learning Techniques

Deep learning is a technique that learns various systems to perform various activities naturally by humans: learn by example. It performs classification tasks by images, text, or sound in deep learning. It achieves the state of accuracy sometimes by exceeding the human-level performance. A set of labeled data and neural network architectures containing many layers are used in the training process. Most deep learning methods use neural network architectures, so deep learning models are often referred to as deep neural networks.

PROBLEM STATEMENT:

Social Media is a valuable communication tool with others locally and worldwide, as well as to share, creates, and spread information. On the other, people with depression might withdraw from face-to-face interactions and spend more time online. Various studies found that people who spend more time on social media tends to have depression.

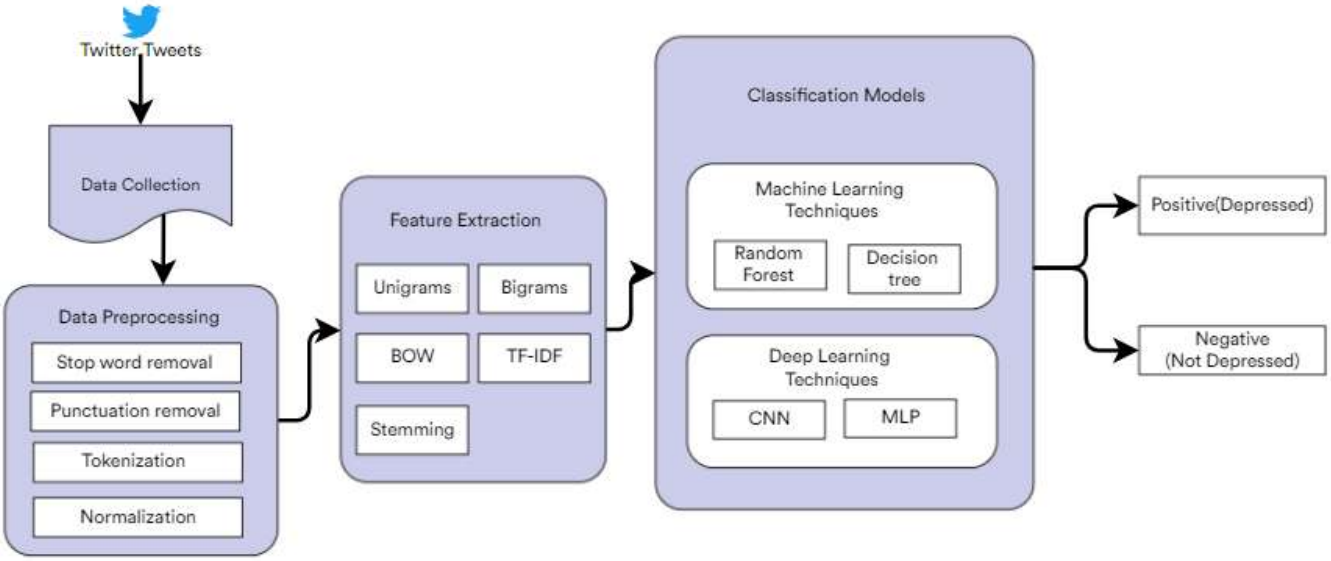
This project aims to detect whether the user is depressed or not using various Machine Learning and Deep Learning.

Many combinations of feature extraction techniques, Machine Learning and Deep Learning classifications are used to improve the accuracy of the prediction.

OBJECTIVE:

To detect a probable depressed Twitter user based on his/her tweets. Machine Learning, Deep Learning and Natural Language Techniques are used for classification process and predicting whether the user is depressed or not.

ARCHITECTURE DIAGRAM:



ARCHITECTURE EXPLANATION:

Data Collection:

The data (twitter tweets) are scraped from the users site. Scrapy is a Python framework for large scale web scraping. It gives you all the tools you need to efficiently extract data from websites, process them as you want, and store them in your preferred structure and format.

Data Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning and deep learning model. It is the first and crucial step while creating a machine learning and deep learning model.

- **Stop word Removal** - Removing the stop word (.) that doesn't have meaning for analysis.
- **Punctuations Removal-** Removing unwanted punctuations.
- **Tokenization-** It is the process of dividing text into a set of meaningful pieces. These pieces are called tokens.

- **Normalization-** It means transforming the data, namely converting the source data in to another format that allows processing data effectively. The main purpose of data normalization is to minimize or even exclude duplicated data.

Feature Extraction:

Feature extraction is the method of selecting and combining data into features reducing the data amount that must be accurately processed and the original data set described thoroughly. The amount of redundant data for a given analysis will also get reduced.

- **Unigram**- Probably the simplest and the most commonly used features for text classification is the presence of single words or tokens in the text. Single words from the training dataset are extracted and frequency distribution is preformed.
- **Bigram**- Two words from the training dataset are extracted and frequency distribution is preformed.
- **BOW**- In this technique, a text is represented as the bag of its words, disregarding grammar and even word order but keeping multiplicity.
- **TF-IDF** – It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- **Stemming**- Stemming is the process of reducing a word to its root form. This ensures variants of a word match during a search. For example, walking and walked can be stemmed to the same root word: walk.

Classification Model:

Machine Learning Classifications:

- **Random Forest-** Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model
- **Decision Tree** - A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

Deep Learning Classifications:

- **CNN-** Convolutional Neural Networks or CNNs are a type of neural networks which involve layers called convolution layers which can interpret spacial data. A convolution layers has a number of filters or kernels which it learns to extract specific types of features from the data. The kernel is a 2D window which is slided over the input data performing the convolution operation. We use temporal convolution in our experiments which is suitable for analyzing sequential data like tweets.
- **MLP-** MLP or Multilayer perceptron is a class of feed-forward neural networks, which has at least three layers of neurons. Each neuron uses a non-linear activation function, and learns with supervision using backpropagation algorithm. It performs well in complex classification problems such as sentiment analysis by learning non-linear models.

LIST OF MODULES:

- Data Extraction and Data Preprocessing.
- Feature Extraction.
- Classification Model.

BRIEF DESCRIPTION OF MODULES:

- **Data Extraction and Data Preprocessing:**
  The Twitter tweets are scraped from the user's site using Python framework Scrapy.
  The extracted data needs to be cleaned to perform feature extraction.
  Stop word, punctuation are removed, then the sentences are tokenized and Normalized.

- **Feature Extraction:**
  It reduces to the processing groups from the initial raw data.
  Many feature extraction techniques are performed to see which techniques gives more accuracy, the list of techniques used are:
  1. Unigram
  2. Bigram
  3. BOW
  4. TF-IDF
  5. Stemming

- **Classification Model:**

  A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.

  Machine Learning Classification Models:

  1. Random Forest
  2. Decision Tree

  Deep Learning Classification Models:

  1. CNN
  2. MLP

REFERENCES:

[1] Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review, SN Computer Science, Published: 19 November 2021, Article number: 74 (2022).

[2] Twitter Sentiment Analysis using Deep Learning , Researchgate, Published: June 13, 2021, DOI: 10.5281/zenodo.5059877