Detection of Depression Using Machine Learning Algorithms

https://doi.org/10.3991/ijoe.v18i04.29051

M Ravi Kumar, Kadoori Pooja, Meghana Udathu, J Lakshmi Prasanna,
Chella Santhosh(✉)
Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India
csanthosh@kluniversity.in

Abstract—Online media outlets such as Facebook, Twitter, and Instagram have forever altered our reality. People are now more connected than ever before, and they have developed such a sophisticated identity. According to ongoing research, there is a link between excessive usage of social media and depression. A mood illness is known as depression. It's defined as sadness, loss, or anger that interferes with a person's day-to-day activity. For different people, depression expresses itself in a number of ways. It might cause disturbances in your daily routine, resulting in missed time and lower productivity. It can also affect relationships as well as some chronic conditions. It has evolved into a serious disease in our generation, with the number of those affected increasing by the day. Some people, on the other hand, can confess that they are depressed, while others are utterly ignorant. On the other hand, the great majority Social media has evolved into a "diary," allowing them to share their mental condition.

Keywords—natural language processing, Naives bayes, logistic regression

1 Introduction

The expansion of internet and communication technologies, particularly online social networks, has revitalized people's electronic interactions and communication. Facebook, Twitter, Instagram, and other social media platforms not only hold textual and multimedia information, but also allow users to express their feelings, emotions, and sentiments about a topic, subject, or issue online. On the one hand, this is excellent for users of social networking sites to openly and freely share and comment to any issue online; on the other hand, it allows health professionals to gain insight into what might be going on in the mind of someone who replied to a topic in a particular way. Machine learning techniques could potentially offer some unique features that can assist in examining the unique patterns hidden in online communication and processing them to reveal the mental state (such as "happiness," "sadness," "anger," "anxiety," and "depression") among social network users to provide such insight. Furthermore, there is a growing corpus of studies addressing the significance of social networks in the form of social interactions such as breakup relationships, mental illness ('depression,''anxiety,'bipolar,'etc.), smoking and drinking relapse, sexual harassment, and suicidal thoughts [1,2].

1.1 Literature survey

This would include a review of depression, existing depression detection systems that employ a variety of ML algorithms and known research gaps that will be addressed in the proposed project. There are several various aspects in this area. Depression is a mental health disorder that has grown in popularity as a topic of conversation in the context of everyday health concerns [3]. A large number of people suffer from the negative impacts of depression, yet only a small percentage receives adequate therapy each year. They also investigated the idea of using social media to detect and assess any signs of serious depression in people. They measured behavioral credits associated with social engagement, feeling, dialect and semantic styles, sense of the self-system, and mentions of antidepressant drugs through their web-based social networking postings. Ignoring depression symptoms or neglecting to treat depression can have serious implications that put one's life in jeopardy [4]. Depression is produced by a complex mix of social, biological, and psychological factors in its early stages. Depression can be caused by a variety of serious and complex conditions. Clinical depression and bipolar disease are the two most frequent types of depression, with clinical depression and bipolar illness being the most common [5]. On the basis of dataset exploration machine learning aids in discovering fascinating patterns and information. Previous depression research has relied on publicly accessible social media data. The researchers used emotional and linguistic forms of word usage to conduct their research. The classification was also carried out using the SVM technique with various kernels [6].

2 Proposed methodology

2.1 Work-flow

- Step 1: In this project the dataset is feed to the model using pandas library.
- Step 2: The Data visualization is done by using the library called seaborn.
- Step 3: we have used re-A regular expression (or RE) describes a collection of strings that match it; the methods in this module let you to verify whether a given text fits a regular expression (or if a given regular expression matches a particular string, which comes down to the same thing).
- Step 4: Trained the algorithm using NLTK library,The model will understand the input data given and respond accordingly.
- Step 5: By using wordcloud library, the model is trained with depression words and not depressed words.
- Step 6: Applying Logistic regression and Naives bayes multinomial model and checking the accuracy.
- Step 7: Checking the output by giving the input data to the model.

2.2 Project execution

This project was completed in JSON (JavaScript Object Notation). It's an open standard file format and data exchange format that stores and transmits data objects made up of attribute–value pairs and array data types using human-readable language.

2.3 Procedure

- Open command promt (windows+r and type cmd and enter).
- Type jupyter notebook and enter.
- Then jupyter notebook will be opened in the respective browser.
- Select the file of data and upload it, then click on new, open python3.
- Load the libraries and data and execute the code as per the commands.

2.4 Libraries

NLTK. Natural Language Toolkit is a collection of natural language tools. It's a collection of statistical language processing libraries and applications. It's one of the most powerful NLP libraries, featuring packages for teaching robots to comprehend human language and respond appropriately [7].

Word cloud. You've probably seen a cloud packed with several words of varying sizes that signify the frequency or significance of each word. The Tag Cloud or Word Cloud is what this is termed. The magnitude of each word represents its frequency or relevance in a word cloud, which is a data visualization tool for visualizing text data. A word cloud can be used to emphasize important textual data points. Data from social networking websites is frequently analyzed using word clouds [8].

Pandas. Pandas is a tool for manipulating large amounts of data at a high level. It's based on the NumPy package. The Data Frame is its primary data structure. Data Frames are a type of tabular data that may be stored and manipulated in rows of observations and columns of variables. It includes data structures and methods for manipulating numerical tables and time series, in particular. Pandas is a widely used open source Python library for data science, data analysis, and machine learning activities. It is developed on top of Numpy, a library that supports multi-dimensional arrays [9].

Seaborn. Seaborn is a Python data visualization tool based on the matplotlib library. It includes a high-level interface for building aesthetically appealing and educational data visualizations. Data frames and the Pandas library are simple to use with Seaborn. The graphs that have been made can be readily altered [10].

Regular expression. The methods in this module allow you to check whether a supplied text matches a regular expression(re) (or if a given regular expression matches a particular string, which comes down to the same thing). A regular expression is a particular sequence of characters that uses a specialized syntax to help you match or locate other strings or collections of strings. In the UNIX realm, regular expressions are commonly employed [11].

2.5 Algorithms

Logistic Regression Model (LRM)

- Type of analysis can help you predict the likelihood of an event happening or a choice being made.
- Logistic model is used to model the probability of a certain class event existing such as win/lose or healthy/sick.
- Supervised learning classification algorithm used to predict the probability of a target variable. Figure 1 shows the hysteresis curve of LRM.

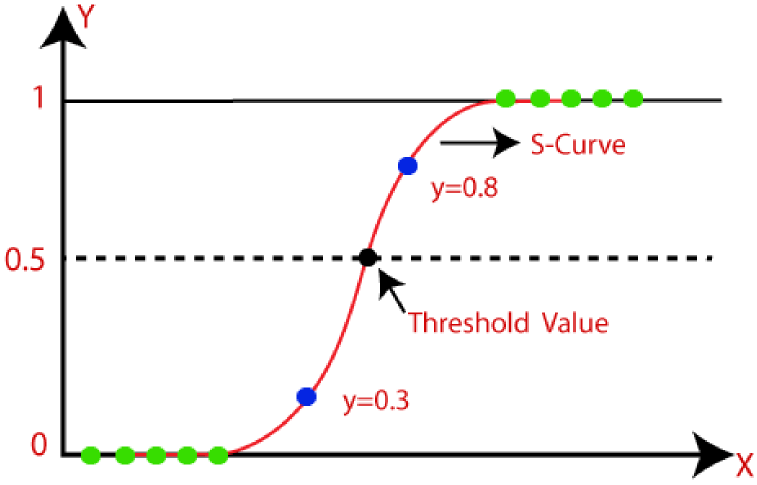


Fig. 1. Logistic regression model hysteresis loop

Naive bayes model

- Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast ML models that can make quick predictions.
- Naive Bayes classification is a form of supervised learning.
- It was initially introduced for text categorization tasks.
- wide variety of classification tasks like sentiment prediction.
- It is easy and fast to predict the class of the test data set. It also performs well in multi-class prediction.

Naives bayes multinomial model. Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature.

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \tag{1}$$

Where P(c|x) = the posterior probability of the class, C is the target and predictor, and x is attributes.

P(c)= the class's prior probability

P(x|c) = The probability of predictor per class, the class known as the likelihood

P(x) = Predictor's prior Probability.

3 Flow chart

In this study, for the detection and processing of depression data received as Twitter posts, we first concentrated on four types of factors: emotional process, temporal process, language style, and all (emotional, temporal, linguistic style) aspects together. We then use supervised machine learning techniques to investigate each factor type separately. 'Decision tree,' 'k-Nearest Neighbor,' 'Support Vector Machine,' and 'ensemble' are considered appropriate classification approaches for each category. The flowchart of work is shown in the Figure 2.

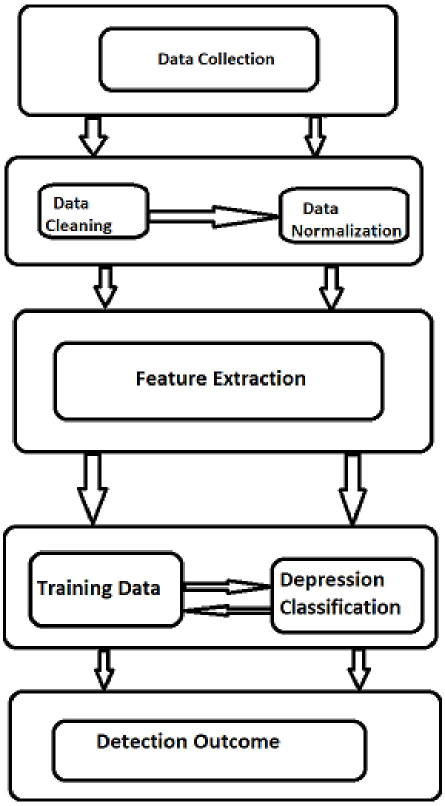
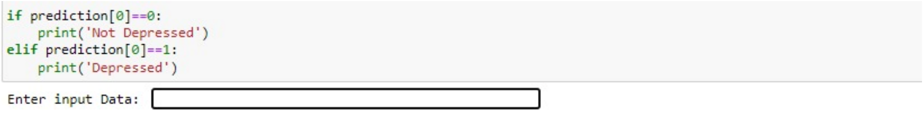


Fig. 2. A methodological overview of Twitter data analysis for depression analysis

4 Result discussion

Input stream:

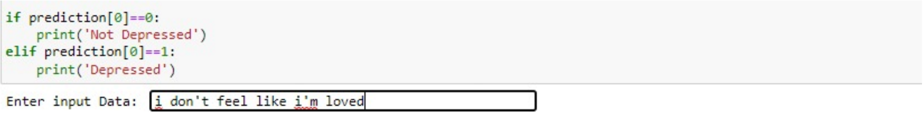


```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data:
```

Case 1: Using the depressed data as input to the model.

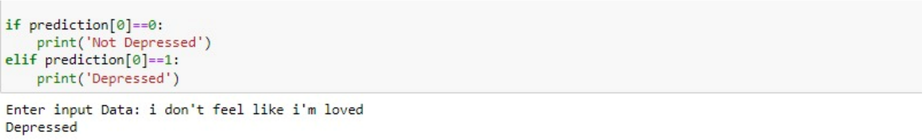
Input (1):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i don't feel like i'm loved
```

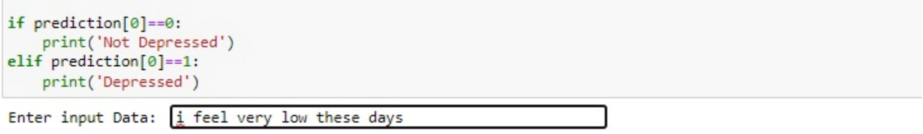
Output (1):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i don't feel like i'm loved
Depressed
```

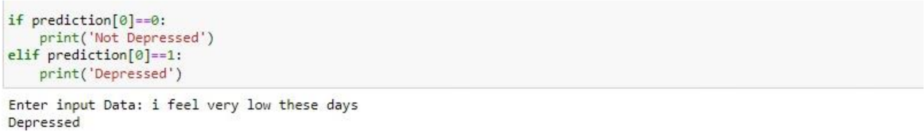
Input (2):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i feel very low these days
```

Output (2):



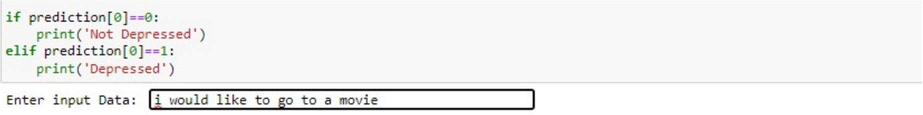
```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i feel very low these days
Depressed
```

As we can observe in the above images, The model analyzed the given input data and predicted the person is depressed, The model is able to predict, whether the person is depressed or not depressed.

Case 2: Using the non-depressed data as input to the model.

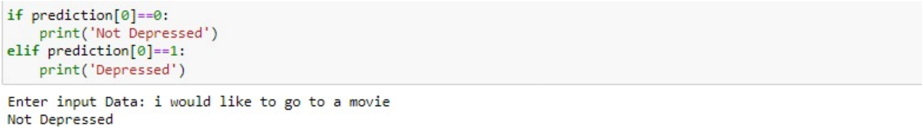
Input (1):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i would like to go to a movie
```

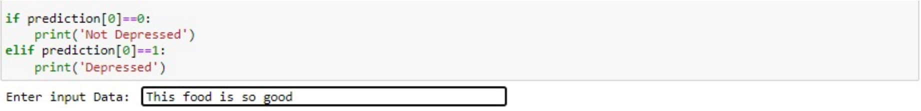
Output (1):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: i would like to go to a movie
Not Depressed
```

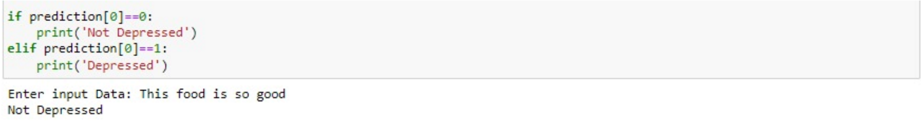
Input (2):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: This food is so good
```

Output (2):



```
if prediction[0]==0:
    print('Not Depressed')
elif prediction[0]==1:
    print('Depressed')
```

```
Enter input Data: This food is so good
Not Depressed
```

As we observe in the above images, the model gave the output as not depressed. The model will predict whether a person is depressed or not depressed based on the input data given to the model.

The output findings for the 1000 tweets dataset that was categorized using the Nave Bayes algorithm [12-15]. The data is accurately classified in 92.34 percent of the cases. The output findings for the 1000 tweets dataset using Logistic Regression classification. The data is accurately classified in 92.34 percent of the cases [16].

The output findings for the dataset that was categorized using the Nave Bayes algorithm. The results demonstrate that 97.31% of the data has been accurately categorized [17]. The output findings for the 3000 tweets dataset using Logistic Regression classification. The results demonstrate that 97.31% of the data has been accurately categorized.

5 Conclusions

In this paper, we demonstrated the ability to use tweeter as a tool for assessing and detecting serious depression among its users in this research. Several research problems were outlined at the beginning of this paper to provide a clear picture of this work. The research issues are revealed by the analytics done on the chosen dataset. The following is a synopsis of our findings: While we all experience mood swings, sadness, or depression from time to time, few people experience these feelings on a regular basis, for long periods of time (weeks, months, or even years), and for no obvious reason. Despondency is more than just a bad mood—a it's real illness that affects a person's bodily and mental well-being. Depression can strike anyone at any time. Some periods or situations, on the other hand, render us more sensitive to depression. Growing older, losing a loved one, starting a family, and retiring can all cause physical and mental changes that might contribute to depression in a small number of people.

6 References

[1] "Depression," World Health Organization. https://www.who.int/health-topics/depression#tab=tab_1

[2] L. Goldman, "Depression: What it is, symptoms, causes, treatment, types, and more," 2019. https://www.medicalnewstoday.com/articles/8933

[3] S. Kemp, "Digital 2020: 3.8 billion people use social media - We Are Social UK - Global Socially-Led Creative Agency," Jan. 30, 2020. https://wearesocial.com/blog/2020/01/digital-2020-3-8- billion-people-use-social-media

[4] Y.-T. Wang, H.-H. Huang, and H.-H. Chen, "A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text."

[5] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," Heal. Inf. Sci. Syst., vol. 6, no. 1, pp. 1–12, 2018. https://doi.org/10.1007/s13755-018-0046-0

[6] A. Sistilli, "Twitter Data Mining: Analyzing Big Data Using Python | Toptal." https://www.toptal.com/python/twitter-datamining-using-python

[7] Fontanella Clint, "How to Get, Use, & Benefit From Twitter's API." https://blog.hubspot.com/website/how-to-use-twitter-api

[8] "Depression." https://www.who.int/news-room/fact-sheets/detail/depression

[9] N. Schimelpfening, "7 Most Common Types of Depression," Aug. 26, 2020. https://www.verywellmind.com/common-typesof-depression-1067313

[10] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni, and A. Ulhaq, "Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique," Int. Conf. Comput. Commun. Chem. Mater. Electron. Eng. IC4ME2 2018, pp. 4–7, 2018. https://doi.org/10.1109/IC4ME2.2018.8465641

[11] H. S. ALSAGRI and M. YKHLEF, "Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features," IEICE Trans. Inf. Syst., vol. E103.D, no. 8, pp. 1825–1832, 2020. https://doi.org/10.1587/transinf.2020EDP7023

[12] M. Nadeem, "Identifying Depression on Twitter," pp. 1–9, 2016.

[13] M. Gaikar, J. Chavan, K. Indore, and R. Shedge, "Depression Detection and Prevention System by Analysing Tweets," SSRN Electron. J., pp. 1–6, 2019. https://doi.org/10.2139/ssrn.3358809

[14] M. M. Aldarwish and H. F. Ahmad, "Predicting Depression Levels Using Social Media Posts," Proc. - 2017 IEEE 13th Int. Symp. Auton. Decentralized Syst. ISADS 2017, pp. 277–280, 2017. https://doi.org/10.1109/ISADS.2017.41

[15] P. Kaviani and S. Dhotre, "International Journal of Advance Engineering and Research Short Survey on Naive Bayes Algorithm," Int. J. Adv. Eng. Res. Dev., vol. 4, no. 11, pp. 607– 611, 2017. https://doi.org/10.21090/IJAERD.40826

[16] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," Proc. 2014 Int. Conf. Data Softw. Eng. ICODSE 2014, no. November 2017, 2014. https://doi.org/10.1109/ICODSE.2014.7062654

[17] S. Narkhede, "Understanding Confusion Matrix | by Sarang Narkhede | Towards Data Science," May 09, 2018. https://towardsdatascience.com/understanding-confusion-matrixa9ad42dcfd62

7 Authors

M Ravi Kumar is Working as Associate Professor in Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh.

Kadoori Pooja is a B. Tech Final Year Student of Electronics and Communications Engineering department, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh.

Meghana Udathu is a B. Tech Final Year Student of Electronics and Communications Engineering department, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh.

J Lakshmi Prasanna is working as Assistant Professor in Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh.

Chella Santhosh is working as Associate Professor in Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh.

Article submitted 2021-12-22. Resubmitted 2022-01-18. Final acceptance 2022-01-19. Final version published as submitted by the authors.