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

Predictive Mental Model for Social Media

Mubarak Sadiqa[1], Pooja Kumari S[2] Raqeeb Ahmed Khan[3] Syeda Afra Maryam[4]
Prof.Sreekantha.b [5] Prof Bibiana Jenifer J[6]

[1,2,3,4] Final year students, [5,6]Assistant Professor Department of Information Science and Engineering,
HKBK College of Engineering , Nagawara ,Bengaluru ,India

Email Id: 1hk18is096@hkbk.edu.in Shreekantha.is@hkbk.edu.in jeniferj.is@hkbk.edu.in

Abstract—**In today's world having a population of 7 Billion, 4 Billion people use Social Media amongst which 300 Million are suffering from depression. Suicide being the second major cause of death among 15-29 years of age group which is often led by depression.**

Mental health in people is a worrying factor and at times it leads to depression and in few cases people may take drastic steps.

The project aim to perform depression analysis on Twitter and Reddit data collected from Social Media by extracting tweets & posts from user's account, preprocessing data and analyzing it by applying sentiment analysis method which is proposed by utilizing vocabulary and man-made dictionary-rules to calculate the sentiment inclination of a person based on his micro-blog.

Thus by using Vader and SIA (Sentiment Intensity Analyzer) we analyze the emotions and intensity of the posts and predict the degree of positive, negative and neutral. Further, applied Naive Bayes classifier consisting of 20 class groups/ target names and 18000 records to predict the group of the post.

Thus, our model aims to provide a unified interface to track and analyse mental health.

Keywords: Vader, SIA, Naïve Bayes Classifier

I. Introduction

Sentiment Analysis is a process of determining whether a 'text' expresses a positive, negative, neutral opinion about the topic. It's far more efficient than manually sorting data because it's completely automated.

Mental health in people is a worrying factor and at times it leads to depression. For those of us who experience mental health concerns, it's often even difficult to explain to our closest friends about how we're feeling but People are unable to generally talk about it because the feelings are so overwhelming, there is a lot of stigma, and friendship groups don't necessarily allow for vulnerable conversations, they allow for fun and gossip but not necessarily vulnerability and because of this a few cases lead to people taking drastic steps.

Social media has become a space where people consistently speak about mental health and caring for others and being there for each other

Social media has a reinforcing nature. Using it activates the brain's reward centre by releasing dopamine, a "feel-good chemical" linked to pleasurable activities such as sex, food, and social interaction. The platforms are designed to be addictive and are associated with anxiety, depression, and even physical ailments.
According to the Pew Research Centre, 69% of adults and 81% of teens in the U.S. use social media. This leads to the large amount of the population is at increased risk of feeling depressed and anxious over their social media use
A 2018 British study tied social media use to decreased, disrupted, and delayed sleep, which is associated with depression, loss of memory, and less academic records. Social media use can affect users' health even more directly. Researchers know the connection between the mind and the gut can turn anxiety and depression into nausea, headaches, muscle tension, and tremors.

Thus the project aim to perform depression analysis on Twitter and Reddit data collected from Social Media.

II. Literature Survey

1. NLP

Natural Language Processing (NLP) is a field of computer science, artificial intelligence, and linguistics which deals with the text data.The goal is for computers to process or "understand" natural language in order to perform various human like tasks like language translation or answering questions.Natural language processing (NLP) techniques can be used to make inferences about peoples' mental states from what they write on Twitter,Reddit and other social media. These inferences can then be used to create online pathways to guide people to health information and assistance.

2. Survey on Sentiment Analysis

Sentiment Analysis is a sub-field of NLP that tries to identify and extract opinions within a given text across blogs, reviews, social media, etc. Sentiment Analysis can help craft all this growing unstructured text into structured data using NLP and open source tools. It is a text analysis method that detects polarity within the text, whether a whole document, paragraph or sentence. It aims to measure the attitude, sentiments, evaluations, attitudes, and emotions of a person based on the computational treatment of subjectivity in a text. For example, Twitter is a platform for expressing people's opinions, thoughts via text know as tweets.

3. *VADER*

VADER (Valence Aware Dictionary for Sentiment Reasoning) is a model used for text sentiment analysis that is about both polarity (positive/negative) and intensity (strength) of emotion. It is available in the NLTK package and can be applied to unlabelled text data.

VADER sentimental analysis uses a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text is obtained by summing up the intensity of each word. For example- Words like ‘love’, ‘fun’, ‘happy’, ‘like’ all convey a positive sentiment. Also, VADER is intelligent enough to understand the basic context of these words, such as “did not love” as a negative statement. It also understands the importance of capitalization and punctuation, such as “ENJOY”

VADER not only tells about the Positivity and Negativity score but also tells us about how positive or negative a sentiment is using compound score. The Compound score is a metric which use to calculate the sum of all the lexicons which have been normalized between -1 (extreme negative) and +1 extreme positive).

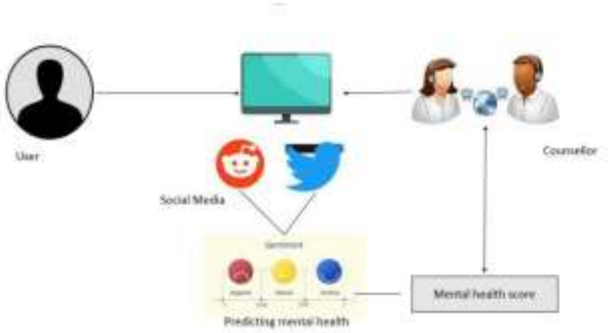
Positive sentiment : IF (compound score >= 0.05) Neutral sentiment : IF (compound score > -0.05) and (compound score < 0.05) Negative sentiment : IF (compound score <= -0.05)

4. *Survey on Social Media*

Social media networks have been developed as a great point for its users to communicate with their interested friends and share their opinions, photos, and videos reflecting their moods, feelings and sentiments. This helps to find an opportunity to analyze social network data for user’s feelings and sentiments to investigate their moods and attitudes when they are communicating via these online tools. Methods diagnosis of depression using social networks data has picked an established position globally, there are several dimensions that are yet to be detected.

III. Proposed Methodology

The proposed system aims to design a Machine Learning(ML) algorithm that works for both the systems (Twitter & reddit) and gives accurate results to predicts the scenario. The implementation was dealt in a way of Divide & Conquer as we had different social media platforms taken into consideration. Each was approached the problem in their own way then later on it was integrated into one single project. For Twitter, most recent tweets were fetched from user’s account and uses sentiment analysis algorithm to predict results. For Reddit, the subreddits were scarped based on the searched keyword to know the thought process and also to understand the sentiment tag attach with each comments. It helps understand the sentiment inclination of a person posting on a similar subreddit.



Twitter

For this process to perform, the Counsellor logins to the dashboard and have access to Twitter and Reddit’s Developer Profile via API. The counselor gets access to a user Twitter account’s most recent Tweets, Retweets, replies similar to what we see on user profile timeline. Then text preprocessing is done followed by sentiment analysis via Natural Processing Toolkit’s VADER and assign compound score of the tweets to know users sentiment inclination.

3.1 Extracting a user’s Tweet

The user tweet timeline endpoints gives us the access to tweets published by a specific twitter user.The user tweet timeline are the Twitter API V2 version endpoints.

This endpoint gives you access to a single Twitter account’s most recent Tweets, Retweets, replies similar to what you can see on user profile timeline.

Reddit

Reddit, the subreddits are scraped based on the searched keyword. The fetched data is then pre-processed followed by sentiment analysis using Natural Processing Toolkit’s VADER which predicts the sentiment of a person by extracting lexicons from his posts and comments and assigns a percentage score of positive, negative and neutral based on semantic orientation. The data that is cleaned and processes are stored in data frames having correlation b/w the post and scores and projected as an overall compound score of the people.

3.2 Reddit API

The use of PRAW (full form)API wrapper helps get access to Reddit’s Developer profile, thus making it easy for web-scraping, getting posts and comments from subreddits etc. This means that the information on some subreddits can be quite valuable. Either for marketing analysis, sentimental analysis or just for archival purposes.

A. *Equations*

4.1 Applying Naive Bayes Classifier

Naive Bayes classifiers are the collection of classification algorithms which are based on Bayes' Theorem. It consists of a family of algorithms and all share a common principle, which is that every pair of features being classified is not dependent on each other.

Naive Bayes classifiers are popularly used for text classification and text analysis.

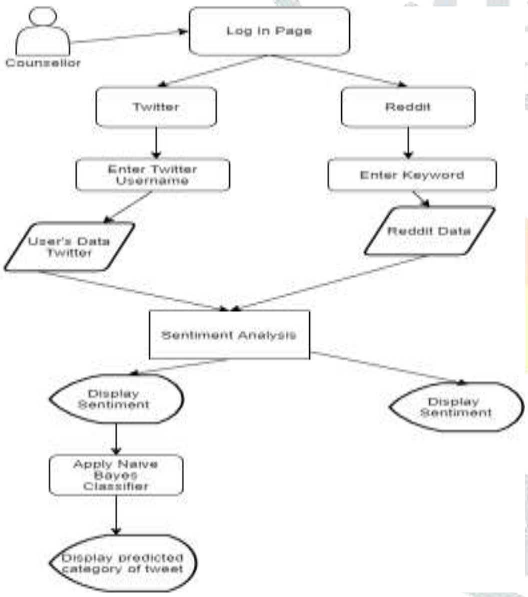
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

(Note: Formula to be written manually)

A, B = events
$P(A \mid B)$ = probability of A given B is true
$P(B \mid A)$ = probability of B given A is true
$P(A), P(B)$ = the independent probabilities of A and B

Figure created by the author

B. *Design of the model*



The Counsellor is presented with a dashboard

Where he/she can login.

The next page is presented with 2 buttons

1) *Twitter Analysis*
2) *Reddit Analysis*

Twitter Analysis

- By clicking on twitter counsellor is provided a field where the username of the targeted user is entered.
- The V2 Endpoint API is called at the backend where the js on file of the user is generated.
- It then extracts the tweets from the js on file and preprocesses it.
- After this the pure text is displayed at frontend using Html/Css which is readable by counsellor.
- We then have a html button which on clicking does sentiment analysis on each tweet of the targeted user and displays the degree of positivity, Negativity and Neutrality of a tweet as we had claimed in our objectives.
- Then we get an option to apply Naive Bayes classifier by clicking the apply button the predicted categories of the tweets are displayed.

Reddit Analysis

- The first page introduces to the panel which takes 'keyword' as an input based on which the extraction of the subreddits are done.

- The PRAW API is called at the back end which crawls through the Reddit's posts, comments and sub-reddits.
- With help of vader and lexicon tools it extracts data from the backend preprocesses it based on sentiments dictionary
- The fetched data is then stored in form of Python's data frame.
- Then it is displayed on the frontend using HTML-CSS in form of progress bar.

Dataset Preparation

The model uses naive Bayes Algorithm. The model was trained using a famous dataset called Fetch_20newgroup It has 20 Classes or Target Names and 18000 datasets. The 20 newsgroups dataset comprises around 18000 newsgroups posts on 20 topics split in two subsets: one for training (or development) and the other one for testing (or for performance evaluation). The split between the train and test set is based upon a messages posted before and after a specific date.

The dataset is divided into two parts, feature matrix and the target vector.

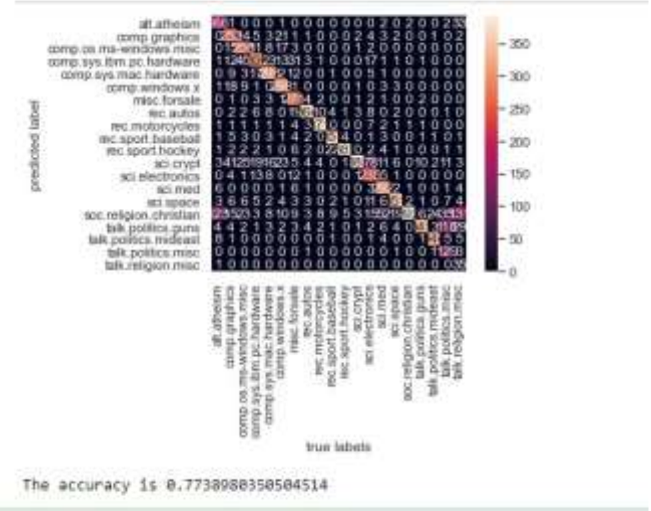
The Feature matrix has all the vectors of the dataset in which each vector consists of the value of dependent features. The number of features is d i.e. X = x1,x2,x2.. xn.

The target vector "Y" contains the value of the group variable for each row of the feature matrix.

We pass the tweets to the trained Model predict the category of tweet.

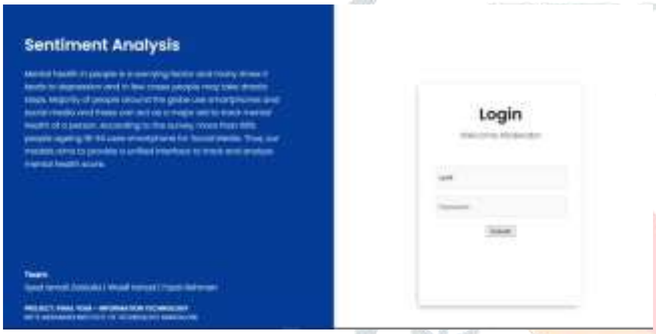
IV. Testing and Results

For our model we have used validation accuracy as an evaluation parameter to test the performance.The accuracy of the model is 77.3 %.Below figure is the confusion matrix of the model.



```
The accuracy is 0.7738980358504514
```

For the output of the Predictive analysis model for mental health each tweet and post is labelled with sentiment score which help the counselor to identify which have negative sentiment inclination.



A login page for the counsellor considering privacy and further the counsellor can check the twitter and reddit results based on the model.
The below figures are the webpages to extract and analyse the tweets and reddit posts from specified user.

A unified interface for the reddit and twitter model which makes it easier for the counsellor to navigate and identify depressed people based on the results from our model.

FUTURE WORK AND SCOPES

In this paper we have exhibited the capability of using VADER and SIA as a tool for measuring and detecting major depression among its users. To give a clear understanding of our work, numbers of research challenges were stated, We found out that social media handles can be a major factor in judging the mental health of a person.

In future work, we plan to use another technique to extract paraphrases from more types of emotional features. Also, we plan to use more dataset to verify our techniques efficiency and effectiveness. We in agreement with the existing body of literature that suggests that more focused studies in depression analysis are needed.

REFERENCES

1) Overview of wireless underground sensor networks for agriculture was demon-strated by Xiaoqing Yu1, Pute Wu1, Wenting Han and Zenglin Zhang. This paper was published in African Journal of Biotechnology Vol. 11(17), pp. 3942-3948, 28 February, 2012.

2) The research of an advanced wireless sensor networks for agriculture demon-strated by Xiaoqing Yu, Pute Wu, Wenting Han and Zenglin Zhang

3) Wireless underground sensor networks, Research challenges demonstrated by Ian F. Akdil z, Erich B. Stentebeck. This paper was published in the year 3 july 2016.

4) Deployment of sensor nodes in precision agriculture using wireless sensor net-work demonstrated by Mrs.Tejal K. Joshi1, Prof.Chirag H. Bhatt, Prof.Tejas R. Kadiya. This paper was published in International Journal For Technological Re-search In Engineering Volume 3, Issue 9, May-2016

5) Wireless Sensor Network deployment for monitoring soil moisture dynamics at the eld scale demonstrated by B. Majonea, F. Vianib, E. Filippic, A. Bellina, A. Massab, G. Tollerd, F. Robolb and M. Saluccib. This paper was published in University of Trento, Department of Information Engineering and Computer Science, Via Sommarive 14, II I 38123 Trento, Italy cPESSL INS c TRUMENTS Italia, Via del Por do 19, II I 38121 Trento, Italy.