**Sentimental Analysis on Social Media to predict Depression**

**Abstract:**

Depression is one of the leading mental health problems and is characterized by persistently low mood, feeling of sadness and loss of interest. Depression diagnosis itself is an active and controversial topic in clinical psychology and psychiatry. Through this research, we have applied Naïve Bayes Classification and NLTK Sentimental Analyzer library on Instagram data to identify the model which can predict the depression more accurately for a user. We have considered the vectorization of user’s status updates using Bag of Words (BOW) and Term Frequency - Inverse Document Frequency (TFIDF). The model will be used to develop an android application which will inform the user about his/her mental status. Based on the prediction and the severity level of depression, the tool will recommend motivational posts or videos for the user suffering from depression. Also, in case of clinical depression it will send a notification to at most 3 close friends of the user so that the person gets the right assistance in time using Android Push Notification library.

**Introduction:**

Social media platforms such as Instagram, Facebook, and Twitter have brought us closer to our family, friends and acquaintances. They have simplified communication even for people living thousands of miles apart from each other. However, the rapidly growing technology has also made people more isolated and depressed than ever before. The advent of social media has created a false sense of connection with others. People are more comfortable in sharing their personal problems and feelings in public with strangers and acquaintances than discussing with family and close friends. Mostly, teenagers and young adult suffer from social media depression. A study published in the Journal of Social and Clinical Psychology empirically confirms the link between social media usage and negative effects on well-being, primarily depression and loneliness.

In this paper, we have proposed a model which will predict the early symptoms of depression and psychiatric disorder in a user. This research will explore the analysis of Instagram users suffering from depression using Naïve Bayes classification method.

**Proposed Model:**

In this model, we have analyzed sentiments from the most recent status updates posted by the Instagram user. We have further classified the data into two different categories.

In first set of data we have analyzed the Instagram hashtags using NLTK Sentimental Analyzer. In the other set, we have analysed comments and status using Multinomial Naïve Bayes.

This dataset is then used to train our model, as it contains large number of words and phrases that can be classified as “positive”, “negative” and “neutral”.

**Case(i): Hashtags analysis using NLTK Sentimental Analyzer**

NLTK is an open source Natural Language Processing (NLP) platform. It is capable of textual tokenisation, parsing, classification, stemming, tagging, semantic reasoning and other computational linguistics. NLTK comes with an inbuilt sentiment analyser module – nltk.sentiment.vader—that can analyse a piece of text and classify the sentences under positive, negative and neutral polarity of sentiments. In this model, we have normalized the hashtag data using json\_normalizer and then applied nltk.sentiment.vader library to analyse the accuracy of positive, negative and neutral hashtags.

Number of negative tags: 98.0

Number of positive tags: 53.0

Number of neutral tags: 347.0

**Case(ii)*:* Review text using Naïve Bayes Classification**

In this case, we have applied Naïve Bayes algorithm on the status and comments mentioned in the Instagram posts. To get the most accurate results we have performed the following steps to identify which model predicts the result more precisely.

**[1] Exploratory Data Analysis**

Exploratory Data Analysis is a critical process of performing initial investigations on data to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

**[1.1] Data Cleaning**

Here we convert the raw dataset into clean dataset. It is observed that sometimes the data have many duplicate entries. Hence it is necessary to remove duplicates to get unbiased results for the analysis of the data.

**[2] Preprocessing**

After the data is cleaned, preprocessing is performed.In the Preprocessing phase, we have followed the below steps:

1. Remove all the html tags present in the text
2. Remove any punctuations or limited set of special characters such as , or . or # etc.
3. Check if the words are made up of English letters and are not alpha-numeric
4. Check to see if the length of the word is greater than 2 (as there are no adjective with 2-letters)
5. Convert all the words to lowercase
6. Remove all the stop words
7. Finally Snowball Stemming the word (as it’s better than Porter Stemming)

Consider the below example on how the output would be after Preprocessing.

*Input: Some days are just bad days, that's all. You have to experience sadness to know happiness, and I remind myself that not every day is going to be a good day, that's just the way it is! 💔*

*Preprocessed Review: days bad days experience sadness know happiness remind not every day going good day way*

After which we collect the words used to describe positive and negative reviews

**[3] Apply Naïve Bayes**

1. **Apply Multinomial Naïve Bayes on these feature sets**

* Convert preprocessed review text into vectors using BOW and TFIDF respectively.

1. **Hyperparamater Tuning (Find best Alpha)**

* Find the best hyper parameter which will give the maximum AUC (Area under the Curve) value
* Consider a wide range of alpha values for hyperparameter tuning, start as low as 0.00001
* Find the best hyper parameter using k-fold cross validation or simple cross validation data
* Use gridsearch cv or randomsearch cv for hyperparameter tuning

1. **Feature importance**

* Find the top 10 features of positive class and top 10 features of negative class for both feature sets and using absolute values of “coef\_” parameter of MultinomialNB and print their corresponding feature names

1. **Representation of results**

* Plot the performance of model both on train data and cross validation data for each hyper parameter, like shown in the figure. Here on X-axis we will have alpha values, since they have a wide range, just to represent those alpha values on the graph, apply log function on those alpha values.
* After finding the best hyper parameter, train the model with it, and find the AUC on test data and plot the ROC curve on both train and test.
* Along with plotting ROC curve, print the confusion matrix with predicted and original labels of test data points. Visualize the confusion matrices using seaborn heatmaps.

1. **Conclusion**

Summarized the result in the table format using the *PrettyTable* library.

|  |  |  |  |
| --- | --- | --- | --- |
| Vectorizer | Model | Hyperparameter alpha | AUC |
| BOW | Multinomial NavieBayes | 10.0 | 0.7374124922401747 |
| TFIDF | Multinomial NaiveBayes | 100.0 | 0.8863648002596138 |

**Conclusion & Future Work:**

In conclusion, the development of automated and objective assessment methods will be valuable for both research and clinical practice. We have proposed Naïve Bayes algorithm with NLTK Sentimental Analyser to provide a number of insights with consistent results, while also identifies many questions open to further investigation. In future, we will further refine our approach to improve the accuracy of our model by applying other algorithms such as logistic regression. We will also consider image recognition libraries to understand the content and sentiment of the photographic data from Instagram. These image classifiers can then be used to predict the mood or personality of the user. This will make our prediction model more robust and accurate.