|  |
| --- |
| Computer script on a screen |
| **STRESS ANALYSIS IN SOCIAL MEDIA**  10868218 |
| |  |  |  | | --- | --- | --- | | Emmanuel Kodjo Djangmah | 8/7/23 | DCIT 316 | |

Department Of Computer Science – University of Ghana

School Of Physical and Mathematical Sciences

DCIT 318 – Computational Models for Social Media Mining

Mr. Mark Attah Mensah

August 7, 2023

Table of Contents

[**ACKNOWLEDGEMENT** 3](#_Toc142373851)

[**INTRODUCTION** 4](#_Toc142373852)

[**PROBLEM** **STATEMENT** 4](#_Toc142373853)

[**MOTIVATION** 4](#_Toc142373854)

[**DATASET** 4](#_Toc142373855)

[**FEATURES** 4](#_Toc142373856)

[**APPROACH** 4](#_Toc142373857)

[**SOLUTION** 5](#_Toc142373858)

[**CONCLUSION** 6](#_Toc142373859)

**INTRODUCTION**

In a globally advancing world and village as we are now, with social media as a room for people from around the globe inconsiderate of distance and origin, social media has become a place of freedom of expression where everyone have a say in any issue of interest as it pleases the user.

Social media has various platforms of interaction, a friendly zone of living, a companion to many in a lonely world as they may be. Being able to reach love ones, meet new people and have relationship with strangers from around the world in a way that is appreciated and enjoyed. Social media is a mass of information bank you can assess no matter what you search for with just a few clicks. Even though helpful, social media can be a place where people can express suicidal thoughts, by means of exposure, spamming from the abuse and stress from other users on one’s life, which can also be as a result of personal experiences. Nevertheless, it is also a safe haven where people who are going through stress, depression, anxiety and various down times can find help and freely express themselves to people and in platforms where they believe they can find help and solution to their problems.

With a rapidly increasing user population of 4.2 billion active users and counting, with the help and presence of web user friendly accessories, many have developed a liking and in some cases addiction to freely discuss their concerns, thoughts and ideas over the internet. Contrarily inflating the user generated content and self-opinionated data. This increases the availability of useful data relating to particular topics of interest over time(s).

Naming a few social media platforms such as Twitter, Instagram, Facebook, Reddit, and Telegram. Through the interactions of its users, the technological and scientific community can lay hands on massive amounts of data, providing the grounds and firm foundation to analyze and study the stress and factors of worries of people around the world(1). Stress is nearly a worldly occurrence; it can be considered as a disease yet the causing factors of stress may be divergent including working activities, abuses, domestic violence, over thinking, bullying of any kind just to mention a few. Stress is a mechanism set in motion by our minds to tackle demanding situations or change at a given time. Although stress is at times helpful, like serving as a source of motivation to tackle demanding situations, excessive stress can lead to a lot of health issues, both physically and mentally. When stress is not addressed at its early stages it can lead to unlikely adverse health problems including psychological disorders, and many more. Some people get rid of stress through taking restful and relaxing naps or sleeps, at times medical aids and some too find it relieving to talk about it to ears that find interest and will pay attention to what they have to say. This is where most people make use of the internet and social media.

In this analysis, we will be building a model to detect the presence of stress in individuals based on their texts on the Reddit platform, as well as the categories of stress such as domestic violence, stress, homelessness, anxiety, assistance, Post-Traumatic Stress Disorder (PTSD), relationship and many more. The significance of this analytical work is to set the stepping stone for future works to develop systems and administrations to help respond to stress texts appropriately.

On a platform such as Reddit, where people can join groups of their choice pertaining to issues of concern in which they believe to find the love, care and attention they so crave for which is healing to their souls and relieving them of their stress and battling worries.

**DATASET.**

The topics under which discussions are made on the reddit platform in various groups and communities are called subreddit. The dataset used in this project is a reddit dataset obtained from Kaggle. Reddit dataset is also preferred in this study because the platform compared to other social media platforms like Twitter allows for longer text write-ups which improves the validity in deducing the complete and whole intension behind any text of a user leading or resulting in accurate predictions of stress.

In this paper we will be dealing with ten subreddit namely:

* domestic violence
* survivors of abuse
* anxiety
* stress
* almost homeless
* homeless
* assistance
* food\_pantry
* ptsd
* relationship

The above subreddit mentioned are the relatively categories we are considering in this paper, as the factors and causes, also the areas and potential bases of stress.

Let’s take a case study of abuse;

Text example “In his own way, I know he loves me but he is double my body weight, he is a weight lifter and he has blind rage that only comes out on me of all people. If I keep gambling, he could permanently damage me. I’m in health care, I know this numbers so why? I like to believe that he knows his limit when we are fighting, but he has scared me and himself in the past. Now that chokings are happening during every incident, the accident could be irreversible.”

We can deduse from the text that (blind rage, permanently damage me, chokings are happening, accident could be irreversible) shows the person is a victim of abuse, this implies that the person is stressing over the issue at hand.

The total data entries in our dataset sums up to 3553 and also 117 features. Table 1shows the distribution of the data entries in their respective sunreddits. It also shows the number of such texts with signs of stress and otherwise.

Table 1:

|  |  |  |  |
| --- | --- | --- | --- |
| **Subreddit** | **Total Post** | **Stress** | **Not Stressed** |
| Anxiety | 650 | 416 | 234 |
| Homeless | 220 | 81 | 139 |
| Domestic violence | 388 | 249 | 139 |
| Assistance | 355 | 126 | 229 |
| Survivors of Abuse | 315 | 143 | 172 |
| Almost Homeless | 99 | 59 | 40 |
| PTSD | 711 | 414 | 297 |
| Stress | 78 | 45 | 33 |
| Food\_Pantry | 43 | 17 | 26 |
| Relationship | 694 | 307 | 387 |
| **TOTAL:** | 3553 | 1857 | 1696 |

METHOD

The dataset obtained from Kaggle contained two .csv files ‘dreaddit-test’ and ‘dreaddit-train’. Upon inspection, it was noticed that both .csv files contained the ‘label’ field which indicates the stress state of text. Due to this, both files were combined in an unordered pattern into one file named ‘data.csv’. The dataset is then later split

During this project, only two features were used in predicting the target variable ‘label’. The features used are ‘subreddit’ and ‘text’. ‘Id’ was not used since it is just a unique identifier for the data entries. ‘Confidence’ was also not used since the project is concerned with whether text signals stress regardless of the confidence level.

The ‘subreddit’ defines the context of the text which was essential in making predictions. Due to this, I combined the ‘text’ and ‘subreddit’ features into ‘text\_sub’ to predict the target variable ‘label’. This decision was made after I found out that combining these two features gives a higher accuracy than working with only the ‘text’ feature.

‘TfidfVectorizer’ was employed in converting the string input into numeric forms to enable learning of the dataset. The binary parameter was set to true which means words would be converted based on their absence or presence signified by 1 and 0 respectively instead of the frequency of their usage which is the case with false. After conversion, ‘Logistic Regression’ is then used to learn the features and make predictions.

SOLUTION

After predictions were made, three metrics were computed to evaluate the competency of the model. The metrics are accuracy, precision and recall and below are the values obtained:

Accuracy: 0.7626

Precision: 0.7485

Recall: 0.8240

The accuracy score reports on the overall performance of the model which is 76%.

The precision score reports on the proportion of the positive predicted stress values that is true which is 74%.

The recall score reports on the proportion of the positive actual stress values that the model correctly predicted which is 82%.

Figure 1 below shows the bar graph distribution of the metrics.

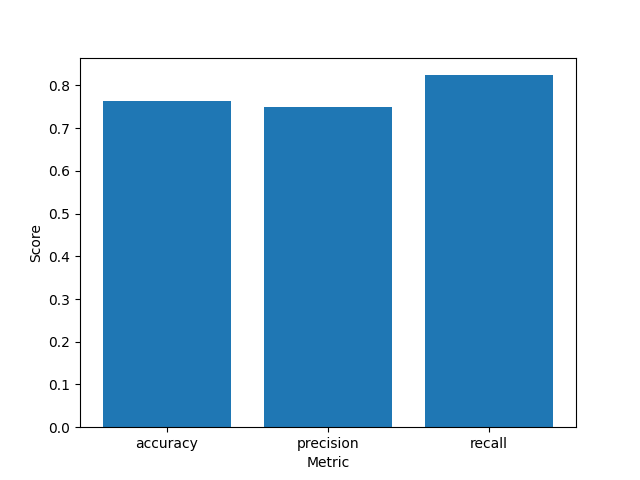


Figure 1

CONCLUSION

With reference to the metric scores obtained, it can be confirmed that the rate at which the model misclassifies stress as non-stress is lower than the rate at which it classifies stress as non-stress. In the health sector, it is better to misclassify a patient without a disease than a patient with a disease. This distinction is very important since the project is health based.

Click [GitHub](https://github.com/Djemm20/10868218_DCIT316_End_Of_Sem_Project.git) to access the repository for this project.