**Stress Analytics in Social Media**

**ABSTRACT**

**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 relative 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 deduct 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.

Let’s take a case study of stress

Text example “I have never really had a problem with my education until this semester. I feel mislead in my classes, i.e. (I do great on the homework and then I end up falling the test). This has led to me falling or doing not up to my standards in my other classes and its beginning to affect my own self-worth instead of being confidence in myself I am resulting to internal hatred and just overall sadness, I am in a relationship with a very loving girlfriend but I feel that if I bring my problems into it, it could go sideways and I don’t want that. Because of all this my sleep schedule has not being exactly ideal, I usually go to bed at midnight and wake up early foe class and the quality of my sleep is not what I will call good (tossing and turning and frequently waking up).

With the above text we can identify or conclude that the person behind the text is stress, example; (I am resulting to internal hatred and just overall sadness, my sleep schedule has not been exactly ideal, falling in class affecting my self-worth.)

Let’s take a case study of ptsd

Text example “I am fine, or so I thought. All my issues with anger, alcohol abuse, frequent bouts of depression, difficulties, difficulties focusing or concentrating at work, I attributed to my bipolar disorder, and that was all I worked on. I have tried so many different types of medications and combinations thereof, more than ten for sure, and even underwent electro-convulsive therapy, but nothing b helped. It was not until these past summer that I thought to try therapy, something I had actively avoided. I didn’t think that talking about anything would affect my bipolar disorder and I really didn’t want to talk about my trauma.”

From the above case we can deduce that the user behind the test is not stressed. Furthermore, the beginning of the conversation seems like he or she was stressed such as “my issues with anger, alcohol abuse, frequent bouts of depression, difficulties focusing or concentrating at work are attributed to my bipolar disorder”. But the later part of the post suggests that the person has find a way to deal with the situation at hand and hence is not stressed but just using the platform to share his or her experience. This is why the reddit platform is preferred in the study, because of the allowance of lengthy text which helps to grasp the whole context of a post. Unlike other social media platforms which have a minimal post length which would in turn affect and handicap the ability to grasp the whole context of post.

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 subreddits. 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**

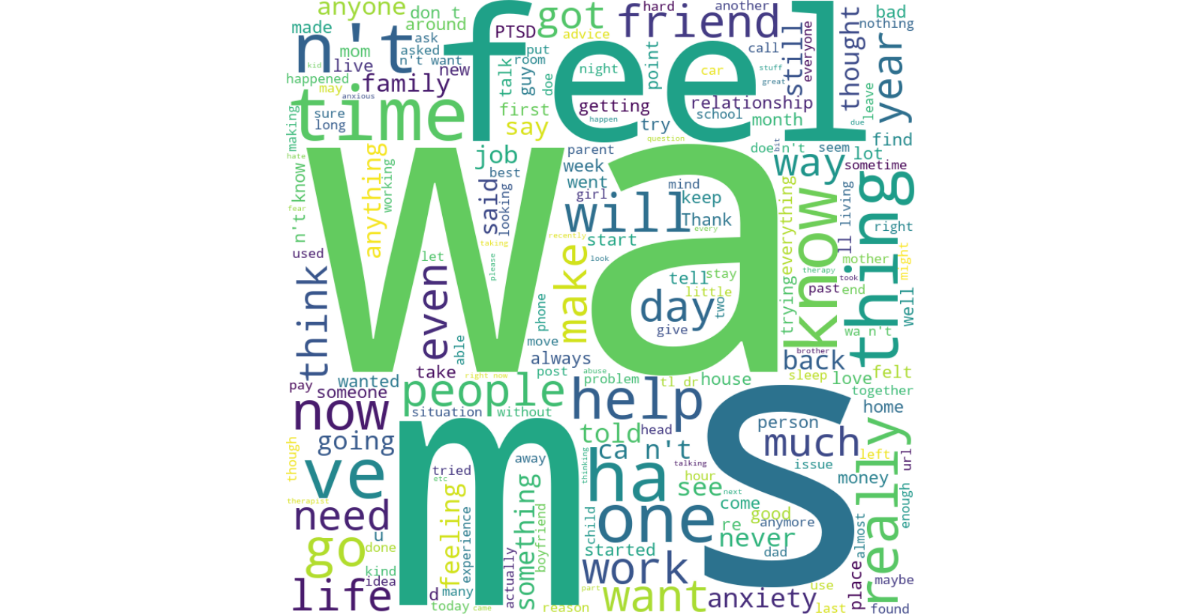


Figure 1 is a word cloud showcasing most frequently used words in the posts. The larger words have higher frequency and usability than the smaller words

Taking a closer at the word cloud shows the popular subreddit under which post are made. These subreddit include anxiety, PTSD, and relationship. This is also confirmed by Table 1 above as it shows that this subreddit have posts of 650 for anxiety, 711 for PTSD and 694 for relationship, which are the highest and top three subreddits under which most people make their posts.

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’.

During this project, only two features were used in predicting the target variable ‘label’. The features used are ‘subreddit’ and ‘text’. The remaining features where not used in the model building process. Two of such features are the ‘Id’ and “confidence”. “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, we combined the ‘text’ and ‘subreddit’ features into ‘text\_sub’ to predict the target variable ‘label’. This decision was made after we found out that combining these two features gives a higher accuracy than working with only the ‘text’ feature.

In order the obtain the best model, we performed three different test. In the three test, we made use of two main classification models; logistic regression and support vector classification (svc). In transforming the individual words to their base form, we made use of ‘WordNetLemmatizer’ and ‘PorterStemmer’. Also for the model to learn the data, we made use of TfidfVectorizer and CountVectorizer to transform the text into a binary form in both scenarios. Below are the three methods in which these libraries were combined in order to determine the best outcome.

Method 1:

The file is then preprocessed by removing special characters, after which we tokenized each post and applied the “WordNetLemmatizer” to resolve the individual words into their basic state. After lemmatization, the tokenized words are joined together to recreate the post and appended to a list. The dataset is then split in code, into train and test sets where the test set is 25% of the dataset. ‘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. Four metrics were computed which are the accuracy, precision, recall and f1 score. Below are the values for each:

* Accuracy: 0.7626546681664792
* Precision: 0.7485380116959064
* Recall: 0.8240343347639485
* F1 score: 0.7844739530132788

Method 2:

The file is then preprocessed by removing special characters, after which we tokenized each post and applied the “PorterStemmer” to resolve the individual words into their basic state. After lemmatization, the tokenized words are joined together to recreate the post and appended to a list. The dataset is then split in code, into train and test sets where the test set is 25% of the dataset. ‘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. Four metrics were computed which are the accuracy, precision, recall and f1 score. Below are the values for each:

* Accuracy: 0.749156355455568
* Precision: 0.7474541751527495
* Recall: 0.7875536480686696
* F1 score: 0.766980146290491

Method 3:

The file is then preprocessed by removing special characters, after which we tokenized each post and applied the “PorterStemmer” to resolve the individual words into their basic state. After lemmatization, the tokenized words are joined together to recreate the post and appended to a list. The dataset is then split in code, into train and test sets where the test set is 25% of the dataset. ‘CountVectorizer 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, ‘SVC’ is then used to learn the features and make predictions. Four metrics were computed which are the accuracy, precision, recall and f1 score. Below are the values for each:

* Accuracy: 0.7277840269966255
* Precision: 0.7285714285714285
* Recall: 0.7660944206008584
* F1 score: 0.7468619246861924

**RESULTS**

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

* Accuracy: 0.7626546681664792
* Precision: 0.7485380116959064
* Recall: 0.8240343347639485
* F1 score: 0.7844739530132788

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%.

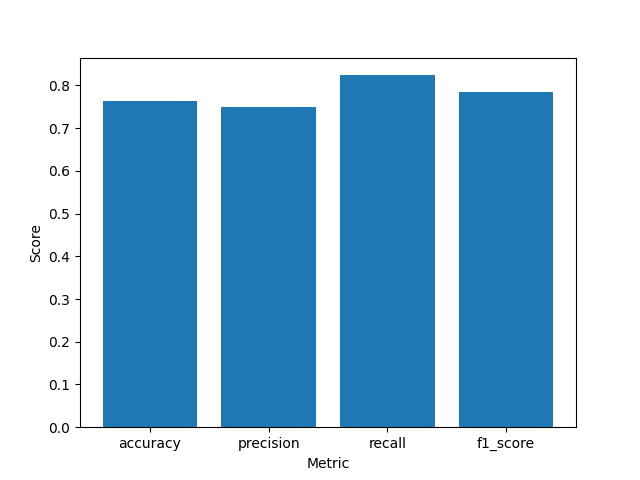


Figure 3: shows the bar graph distribution of the metrics.

**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.