# Applying Machine learning algorithms to detect stressed text in social media comment


| | |
|---|---|
| Jesmin Akhter | 180104052 |
| Aysha Shiddika | 180104054 |
| MD Shafique | 180104059 |


**Project Report**

**Course ID: CSE 4214**

**Course Name: Pattern Recognition Lab**

**Semester: Spring 2021**


**Department of Computer Science and Engineering**

**Ahsanullah University of Science and Technology**

Dhaka, Bangladesh


Sep-04-2022

# Applying Machine learning algorithms to detect stressed text in social media comment


Submitted by

| | |
|---|---|
| **Jesmin Akhter** | **180104052** |
| **Aysha Shiddika** | **180104054** |
| **MD Shafique** | **180104059** |

Submitted To

**Faisal Muhammad Shah**

**Sajib Kumar Saha Joy**,

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology



**Department of Computer Science and Engineering**

**Ahsanullah University of Science and Technology**

Dhaka, Bangladesh

Sep-04-2022

# ABSTRACT

When a mental health issue occurs and its symptoms frequently stress out and impair capacity to operate, it becomes a mental disease.Emotion plays a significant role in daily interpersonal human interactions. This is Necessary to make rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions [1], speech [2], physiological signals [3], etc. Stress detection from the text of social media is one of the most important and difficult tasks in the field of machine learning using natural language processing. Feeling overwhelmed or more emotional can be a mental disorder. Emotion recognition provides benefits to many institutions and aspects of life. It is useful and important for security and healthcare purposes. We wish to work on this topic emotion detection using NLP to address this issue and prevent any undesirable situations from occurring. Natural language processing, machine learning techniques, and neural network architectures to create, tune, and evaluate models that categorize Reedit text comments as "Stressed" or "not-stressed."For this Reedit data set, we got ——accuracy by using machine learning models.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Social networking platforms have become an essential means for communicating feelings to the entire world due to rapid expansion in the Internet era. Several people express their feelings or viewpoints through textual content, pictures, audio, and video. Text communication through Web-based networking media, on the other hand, can be overwhelming. Emotion is one of the most difficult concepts to define in psychology. In fact, there are different definitions of emotions in the scientific literature. Emotion is often involved with temperament, mood, personality, motivation, and behavior. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes. These changes influence thought and behavior. According to other theories, emotions are not causal forces but simply syndromes of components such as feeling, behavior, and physiological changes. The term “emotion” has also been used to describe a highly complex state that is associated with a wide range of mental, physiological, and physical events.

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous). Facial expression is one of them. The human face is extremely expressive, capable of expressing a wide range of emotions without saying a single word. Using physiological signals to recognize emotions can also help people who have been suffering from physical or mental illness for a long time and have difficulty with facial expressions or tone of voice. According to recent research, Emotion detection using natural language processing from the text of one’s post or comments on social media platforms such as Twitter, Facebook, Instagram, and others.

Sentiment analysis or opinion classification has become the buzzword in social network services as people have moved to the digital world to express themselves. A fundamental task in sentiment analysis is determining whether the expressed opinion in a document, sentence, or entity feature/aspect is positive or negative according to the level of the document, sentence, or feature/aspect.

Natural Language Processing (NLP) is a subfield of artificial intelligence (AI). It helps machines to process and understand the human language so that they can automatically perform repetitive tasks.The main objective of this research is to reliably detect positive and negative emotions in English-language tweets on Twitter and comments or posts on Facebook. So, Natural Language Processing (NLP) approaches can be useful here to detect the mental illness. Among all machine learning models, we will use the logistic algorithm,Support Vector Machines,Naive Bayes and the random forest algorithm etc.

We have almost 3554 data. It is good enough to train the models and also we tried in this work

# Chapter 2

# Literature Reviews

There is a substantial body of literature that investigates that determine mental illness reading social media comments.

They reviewed existing techniques for emotion and sentiment detection in this paper [4]. They investigated the performance of machine learning and deep learning algorithms in relation to pre-processing, feature extraction, and dataset size. They also discovered that the lexicon-based technique works well for sentiment and emotion analysis. On the other hand, The dictionary-based approach is highly adaptable and simple to implement, whereas the corpus-based method is based on rules that work well in a specific domain. As a result, while corpus-based approaches are more accurate, they lack generalizability. When the dataset is large, the deep learning approach showed better performance than the machine learning approach. They used Recurrent neural networks, especially the LSTM model, were prevalent in sentiment and emotion analysis, as they can cover long-term dependencies and extract features very well. The lexicon-based approach and machine learning approach (traditional approaches) were both evolving and producing better results in a specific way.

In this Paper [2],Author introduce ,the interaction of features generated from the same audio source was usually ignored, which could result in irrelevant features and increase computational costs. Feature selection method based on correlation analysis and Fisher is proposed, which can remove the redundant features. An emotion recognition method based on an extreme learning machine (ELM) decision tree is proposed based on the confusion degree among different basic emotions to improve the recognition performance of the feature subset after proposal feature selection. Their proposal achieved an average recognition rate of 89.6% . It would be fast and efficient to distinguish different speakers' emotional states from speech, and it would enable future interaction between speaker-independent and computer/robot.

In paper [5] author worked with 4000 data for Sentiment Analysis-Based Sexual Harassment Detection Using Machine Learning Techniques. Here, these 4000 data were collected from ”maps.safecity”. The data were labeled into 3 groups. 1) Ogling 2) Commenting and 3) Gropping. Tokenization, Stemming and Lemmatization used for data pre-processing. The data was split into ratios (80%, 20%) for training and testing after pre-processing the dataset. Tf-Idf was used for feature extraction. 8 different well-known classifiers were run on the same training and testing datasets in order to evaluate the proposed model. Random Forest, Multinomial NB, SVS, Linear SVC, SGD, Bernoulli NB, DT, K Neighbors. The model achieved 81% accuracy using SGD classifier.

In paper [6] author introduced hate speech against women in English Tweets. Dataset with 5000 data were collected from twitter. The detection procedure divided into 2 subtasks. Subtask A - classify whether the data misogyny or not (binary classification). Subtask B - correct multiclass classification problem. NLTK was used to tokenize the tweets and remove English words in both subtask A and subtask B. Bag of Words, Lexical Features and Sentiment Scores were used as Feature Extraction for both subtask A and subtask B. Random Forest, SVM, Gradient Boosting, Linear Regression were used to classify in subtask A. Deep learning and Ensemble learning are used for subtask B. For subtask B, the classification models are: Logistic Regression, Random Forest, Stochastic Gradient Descent (SGD), Multinomial Naive Bayes, XGBOSST, LSTM. Better result found through Deep Learning. 63% accuracy achieved on English Dataset.

In this study [7], five distinct levels of anxiety, depression, and stress intensity were assessed using machine learning algorithms. According to the World Health Organization (WHO), having depression affects more than 300 million people globally and is the most common mental illness. A basic survey evaluating the common signs of stress, sadness, and anxiety was used to gather the data (DASS-21). Following that, Decision Tree (DT), Random Forest Tree (RFT), Naive Bayes, Support Vector Machine (SVM), and K-Nearest Neighbor were used as five distinct categorization algorithms (KNN).The accuracy of naïve Bayes was found to be the highest, although Random Forest was identified as the best model. Due to the fact that this problem produced imbalanced classes, the best-model selection was made on the basis of the f1 score, which is used for cases of imbalanced partitioning. Furthermore, the specificity parameter revealed that the algorithms were also especially sensitive to negative results.Despite Random Forest being the best model, naive Bayes was shown to have the highest accuracy. The better selection was based on the f1 score, which is employed in instances of imbalanced partitioning, because this problem resulted in unbalanced classes. The algorithms were also very sensitive to negative outcomes, as the specificity parameter showed.

# Chapter 3

# Data Collection & Processing

## 3.1 Data Collection

We collected the tweeter post and comments from "kaggle" and the facebook coments with subreddit from the google dataset. Then we merged these texts together and labeled them with 1 and 2 respectively.Dataset has 3554 records and 5 features.This is a Binary Classification Problem.Target Feature is stress.

## 3.2 Data Preprocessing

Our collected data samples were very noisy. We have followed some conventional preprocessing techniques to remove those noise.

- **Removing , @, URL**

  Our text are generated from Facebook,twitter posts and comments.So it has many unnecessary elements in the text. Hashtag, Mentioning someone and URL are one of them. To get the main text we remove ,@ and URL from our dataset.

- **Making all text lowercase**

  In dataset,Upper case is present.For model fitting we convert all text data to lower case.For this purpose we use lower method.

- **Removing punctuation**

  We removed punctuations Like-> ";',!" etc.As they have no contextual meaning in a sentence.

- **Removing Stopwords**

  Especially for our domain stopwords have no worthy use. So we got rid of any kinds of these words.

- **Removing emoji**

  A lot of emojis were there in our dataset. As punctuation emojis don't give any contextual knowledge, we removed the emojis using demoji.

- **Stemming**

  We used SnowballStemmer for stemming our data words. i.e: easily -> easili

- **Lemmatization**

  We used Word Net Lemmatizer for lemmatizing. i.e: changing -> change

# Chapter 4

# Methodology

## 4.1 Data Annotation

For getting accurate annotation, we have classified our dataset into 2 classes. The 2 classes are

- **Stress**
  These data refer to express stress form of mental illness This class is annotated by '1'. E.g. (I loss my job)

- **Not Stress**
  These data refer to express not stress form of mental illness. This class is annotated by '0'. E.g. (I cherished spent the weekend with my buddies.)

Table 4.1: Annotated Data before classification.

| Category | Amount |
|---|---|
| Stress | 1857 |
| Not Stress | 1696 |

We have gathered almost 3554 data, where the number of Stress based text is 1857 and the number of not stress based text is 1696. So, we can say our dataset represents a balanced dataset.

Table 4.2: Annotated Data before classification.

| Raw Text | |
|---|---|
| I am a domestic violence survivor who is still struggling, even after over four years. | domestic violence |
| We were living together life pretty sweet. | |

## 4.2 Word Embedding/ Word Vectorization with TF-IDF

TF-IDF Vectorization is one of the most basic and widely used techniques for converting words to vectors. This approach performs admirably on a smaller dataset. Although it cannot express a word's semantic meaning. In certain circumstances, the sentences rely heavily on a few keywords. As a result, it is an appropriate method for a domain. TF-IDF was built using scikit learn.

### 4.2.1 TF Numerical Value

Term Frequency is abbreviated as TF. Terms refer to the words in a sentence. The frequency of occurrence of that specific term is measured. So, the notion is that if a word appears numerous times in a phrase, it is particularly important to the sentence's construction. The TF approach provides us with such significance as numerical data.

TF(word, document) = count of word in documnet /number of words in the document

### 4.2.2 DF Numerical Value

DF stands for Document Frequency which is similar as TF. The only difference is it counts the occurrences of a word in N documents.

*DF(word) = occurrence of word in N documents*
The actual reason for calculating this is to get the informativeness of a word. DF is the exact inverse of it.

### 4.2.3 IDF Numerical Value

IDF stands for Inverse Document Frequency which gives the informative measures of a word. IDF is typically used to boost the scores of words that are unique to a document with the hope that we surface high information words that characterize our document and suppress words that don't carry much weight in a document. So, in terms of the informativeness of a word, we need relative weightage.

IDF(word) = N / DF(word)

But there is a problem here. N can be very large which is the total number of documents/sentences in the corpus. DF(word) is the number of documents in which the word appears in.

Therefore, we have to scale down the value. Here, Log() comes into the play. Now, we have a new formula.

IDF(word) = log (N/DF(word)+1)

Now, Log() will save us from exploded value of IDF. As we cannot divide anything by 0, we smoothen the value by adding 1 to the denominator.

### 4.2.4 TF-IDF Numerical Value

The required approach is TF-IDF, for which we already determined the TF, DF, and IDF values. The relevance of a phrase is inversely proportional to its frequency across documents, as measured by TF-IDF. TF denotes how frequently a term appears in a document, whereas IDF indicates the relative rarity of a phrase in the collection of documents.

Finally, by taking a multiplicative value of TF and IDF, we get the TF-IDF score.

TF-IDF(word, document) = $TF$(word, document) $\times$ $IDF$(word)

The higher the TF-IDF score the more important or relevant the term is as a term gets less relevant, its TF-IDF score will approach zero.

## 4.3 Bag of Words

The bag-of-words model is a representation that simplifies natural language processing and information retrieval. A text is represented in this paradigm as a bag of its words, ignoring syntax and even word order but retaining multiplicity. It is a text representation that defines the appearance of words within a document. It entails two steps: A list of well-known terms. A metric for the presence of well-known terms.

### 4.3.1 N-Grams

An n-gram is a continuous sequence of n elements from a given sample of text or voice in the disciplines of computational linguistics and probability. Depending on the application, the elements might be phonemes, syllables, letters, words, or base pairs. Typically, n-grams are extracted from a text or audio corpus. N-Grams are —

- **Uni-gram**
  Taking one word at a time for vectorization.
  Example: you are not funny at all
  Uni-gram text: 'you', 'are', 'not', 'funny', 'at', 'all'

- **Bi-gram**
  Taking two consecutive word at a time for vectoization. Example: you are not funny at all Bi-gram text: 'you are', 'are not', 'not funny', 'funny at', 'at all'

- **Tri-gram**
  Taking three words at a time for vectorization. Example: you are not funny at all Tri-gram text: 'you are not', 'are not funny', 'not funny at', 'funny at all'

### 4.3.2 Count Vectorizer

The CountVectorizer technique converts text to numerical data. It will tokenize the input and divide it into n-gram chunks, the length of which we can specify by giving a tuple to the ngram range function. Because the count vectorizer generates a matrix with document and token counts (bag of terms/tokens), it is also known as the document term matrix (dtm).

## 4.4 Machine Learning Models

We have tested the dataset over several models. Such as Logistic Regression, Support Vector Machine(SVM), Random Forest and Naive Bayes. We have picked the best models which preformed quite well. We also tested our dataset with some ensemble learning models such as gradient boosting, adaptive boosting(AdaBoost) and XGBoost.

### 4.4.1 Logistic Regression

The logistic model is a statistical model that represents the chance of one event occurring by making the event's log-odds a linear combination of one or more independent variables. Logistic regression is used in regression analysis to estimate the parameters of a logistic model. Logistic regression is used to address classification difficulties, with binary logistic regression being the most prevalent use (yes or no). Logistic regression is classified into three types: binary, multinomial, and ordinal [8].

### 4.4.2 Support Vector Machine

Support Vector Machines, or SVMs, are used to solve classification and regression problems. However, it is primarily used in Machine Learning for Classification problems. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may simply place fresh data points in the proper category in the future. A hyperplane is the optimal choice boundary. SVM selects the extreme points/vectors that aid in the creation of the hyperplane. These extreme examples are referred to as support vectors, and the method is known as a Support Vector Machine.

### 4.4.3 Radmon Forest

A random forest algorithm [9] is made up of several decision trees. The random forest algorithm's 'forest' is trained via bagging or bootstrap aggregation. The outcome is determined by the (random forest) algorithm based on the predictions of the decision trees. It forecasts by averaging or averaging the output of several trees. The precision of the output improves as the number of trees grows. A random forest method overcomes the constraints of the decision tree technique. It reduces dataset overfitting and improves accuracy. It generates forecasts without requiring extensive package parameters.

### 4.4.4 Naïve Bayes Classifier

The term "Naive Bayes classifiers" refers to a set of classification methods based on Bayes' Theorem. It is a family of algorithms rather than a single method.

share a common premise in that each pair of attributes being categorised is independent of one another

of one another It is mostly employed in high-dimensional text categorization.

dataset for training.It is a probabilistic classifier, which means it predicts based on an object's likelihood. It is dubbed Naive because it believes that the presence of one trait is unrelated to the occurrence of others. Then it is referred to as Bayes since it is based on the idea of Bayes' Theorem. Bayes' theorem, often known as Bayes' rule or Bayes' law, is a mathematical formula used to calculate the probability of a hypothesis given past knowledge. It is determined by the conditional probability.

### 4.4.5 Gaussian Naïve Bayes Classifier

## 4.5 Boosting Models

Boosting models are statistical machine learning models that are combined. For example, Decision Tree, KNN, and so on. Ensemble models are mostly used to improve the performance of statistical models. These ensemble models employ some innovative strategies based on statistical model learning. Furthermore, in imblanced datasets, these ensemble models perform well.

### 4.5.1 XGBoost

XgBoost is an abbreviation for Extreme Gradient Boosting. It is a Gradient Boosted decision tree implementation. This technique generates decision trees in a sequential fashion. Weights are very significant in XGBoost. All of the independent variables are given weights, which are subsequently put into the decision tree, which predicts results. The weight of factors that the tree predicted incorrectly is raised, and these variables are subsequently put into the second decision tree. These various classifiers/predictors are then combined to form a more powerful and precise model. It can solve issues including regression, classification, ranking, and user-defined prediction.

# Chapter 5

# Experiments and Results

## 5.1 Evaluation Matrices Used

- **Accuracy**

  It's the ratio of the correctly labeled subjects to the whole pool of subjects. Accuracy is the most intuitive one. Accuracy answers the following question: How many students did we correctly label out of all the students? Accuracy = (TP+TN)/(TP+FP+FN+TN) numerator: all correctly labeled subject (All trues) denominator: all subjects.

- **Precision**

  Precision is the ratio of the correctly +ve labeled by our program to all +ve labeled. Precision answers the following: How many of those who we labeled as diabetic are actually diabetic? Precision = TP/(TP+FP) numerator: +ve labeled diabetic people. denominator: all +ve labeled by our program (whether they're diabetic or not in reality).

- **F1-Score**
  F1 Score considers both precision and recall. It is the harmonic mean(average) of the precision and recall. F1 Score is best if there is some sort of balance between precision (p) recall (r) in the system. Oppositely F1 Score isn't so high if one measure is improved at the expense of the other. For example, if P is 1 R is 0, F1 score is 0. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

- **Recall**

  Recall is the ratio of the correctly +ve labeled by our program to all who are diabetic in reality. Recall answers the following question: Of all the people who are diabetic, how many of those we correctly predict? Recall = TP/(TP+FN) numerator: +ve labeled diabetic people. denominator: all people who are diabetic (whether detected by our program or not)

## 5.2 Result

## 5.3 Overall Performance

## 5.4 ROC curve and AUC

# Chapter 6

# Future Work and Conclusion

## 6.1 Future Work

Our research purpose proposed a detection of stress types data to detect earlier in social networking sites. In this paper we propoes 7 diffrent machine learning algorithms using BoW and Tf-Idf. We found LR, RF, SVM and Xgboost with Tf-Idf performed very well with our dataset obtaining accuracy of ——96%. Deep learning algorithms and Transformers might be a good option to detect mental illness for large dataset. For doing so, we have to prepare huge amount of customized data and those data need to be pre processed properly. In future, we will try to increase the dataset and apply Deep learning and transformers like: LSTM, Bi-LSTM, BERT, RoBERTA to classify and detect these sort of texts.

## 6.2 Conclusion

Many people experience mental health issues from time to time. When recurring indications and symptoms create regular stress and impair your capacity to operate, a mental health condition becomes a mental disease. A mental condition can be depressing and cause challenges in everyday living. In this paper, we use several forms of machine learning models to determine if someone is stressed or not by analyzing their social media remarks. We use data from Facebook comments and tweeters.

# References


[1] H. Ali, M. Hariharan, S. Yaacob, and A. H. Adom, "Facial emotion recognition using empirical mode decomposition," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1261–1277, 2015.

[2] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, and G.-Z. Tan, "Speech emotion recognition based on feature selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.

[3] M. Ragot, N. Martin, S. Em, N. Pallamin, and J.-M. Diverrez, "Emotion recognition using physiological signals: laboratory vs. wearable sensors," in *International Conference on Applied Human Factors and Ergonomics*, pp. 15–22, Springer, 2017.

[4] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, no. 1, pp. 1–19, 2021.

[5] E. Alawneh, M. Al-Fawa'reh, M. T. Jafar, and M. Al Fayoumi, "Sentiment analysis-based sexual harassment detection using machine learning techniques," in *2021 international symposium on electronics and smart devices (ISESD)*, pp. 1–6, IEEE, 2021.

[6] R. Ahluwalia, H. Soni, E. Callow, A. Nascimento, and M. De Cock, "Detecting hate speech against women in english tweets," *EVALITA Evaluation of NLP and Speech Tools for Italian*, vol. 12, p. 194, 2018.

[7] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258–1267, 2020.

[8] V. Suthar, R. A. Tarmizi, H. Midi, and M. B. Adam, "Students' beliefs on mathematics and achievement of university students: Logistics regression analysis," *Procedia-Social and Behavioral Sciences*, vol. 8, pp. 525–531, 2010.

[9] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar mod-

eling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Saturday 3rd September, 2022 at 7:27am.