Stress Sentimental Analysis Using Machine learning (Reddit): A Review

Sharad Rajyaguru

Student (Pg student), Department of Computer Engineering, Birla Vishvakarma Mahavidyalaya, Anand, Gujarat, India

---***---

Abstract - *Mental health-related issues are one of the leading problems in the world. There are so many health issues that happen due to the extensive usage of social media. In this paper, we survey text mining with help of Reedit dataset. In this survey, paper Researchers applied different pre- processing techniques as well as an applied different feature extraction techniques. Researchers applied preliminary supervised learning approaches for detecting stress, both neural and traditional, as well as an analysis of the data's complexity and diversity, as well as the features of each category.*

Keywords: **Prediction, Stress, data mining, Reedit, dataset, supervised algorithms, NLP**

1. Introduction

A country's well-being is dependent on its citizens' health. A sudden outbreak of a disease can send people into a panic. The importance of early disease identification cannot be overstated. It's crucial to keep it from spreading across a region. The help of the public's use of social media platforms is increasing rapidly. User postings can be of great assistance to those who want to voice their ideas. Keep track of disease outbreaks in different parts of the world. As evidenced by the inclusion of mental health in the Sustainable Development Goals; there has been a growing recognition of the critical role mental health plays in achieving global development goals in recent years. One of the primary causes of disability is depression. Suicide is the second-highest cause of death in people aged 15 to 29. People with serious mental illnesses die considerably sooner than they should – up to two decades earlier – as a result of preventable physical ailments.[1] Even many more peoples are suffering from mental stress according to the survey of TCOH (Delhi-based org) show that more than 74% and 88% of Indians, respectively, experience tension and anxiety.[2]The author collects data sets from one most popular social media forms is Reddit, Reason for selecting this data because there is no text limitation in Reddit data set like Twitter. Reddit data set is open source and free to download. According to a poll report by Marsh India, the country's largest insurance broker, three out of five employees in India (59 percent) reported feeling extremely, oderately, or slightly stressed daily, a higher level than both the worldwide and Asia area norms.

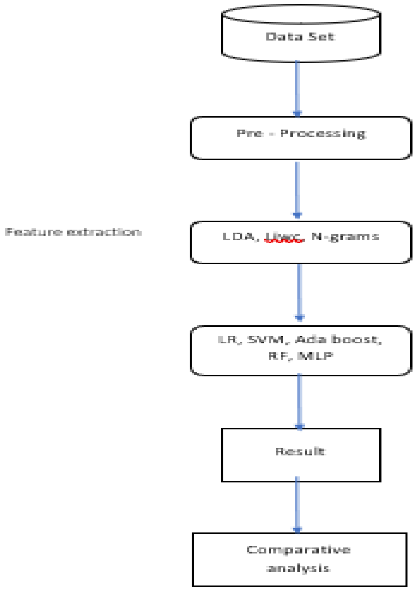
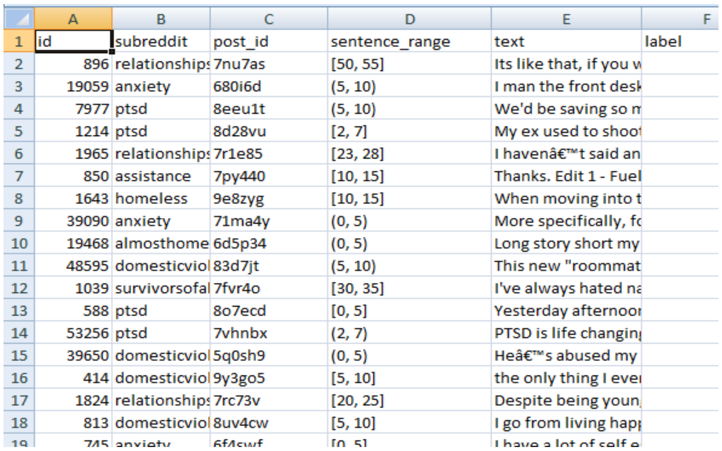


Fig.1. Block Diagram Stress sentiment analysis. [4]

2. Dataset detailed

Reddit is a social media website where user's post-post- topic-specific communities called subreddits. This data was created by Elsbeth Turcan and Kathleen McKeown.[3] These data sets are divided into five categories like abuse, PTSD, anxiety, Social, and Financial. There are both types of Dataset available. Train data set, testing dataset. The size of the Train is 2675 kb and the test data size is 668 kb. Num of raw in 2838 & column is 116 in Training data set and testing data set no raw 716 and column is a 116. There are also many columns available in this dataset.

Label ---> 1 (Stress) Label ---> 0 (Not stress)



	A	B	C	D	E	F
1	id	subreddit	post_id	sentence_range	text	label
2	896	relationships	7nu7as	[50, 55]	Its like that, if you w	
3	19059	anxiety	680i6d	(5, 10)	I man the front desk	
4	7977	ptsd	8eeu1t	(5, 10)	We'd be saving so m	
5	1214	ptsd	8d28vu	[2, 7]	My ex used to shoot	
6	1965	relationships	7r1e85	[23, 28]	I havenâ€™t said an	
7	850	assistance	7py440	[10, 15]	Thanks. Edit 1 - Fuel	
8	1643	homeless	9e8zyg	[10, 15]	When moving into t	
9	39090	anxiety	71ma4y	(0, 5)	More specifically, fo	
10	19468	almosthome	6d5p34	(0, 5)	Long story short my	
11	48595	domesticviol	83d7jt	(5, 10)	This new "roommat	
12	1039	survivorsofal	7fvr4o	[30, 35]	I've always hated na	
13	588	ptsd	8o7ecd	[0, 5]	Yesterday afternoon	
14	53256	ptsd	7vhnbx	(2, 7)	PTSD is life changing	
15	39650	domesticviol	5q0sh9	(0, 5)	Heâ€™s abused my	
16	414	domesticviol	9y3go5	[5, 10]	the only thing I ever	
17	1824	relationships	7rc73v	[20, 25]	Despite being young	
18	813	domesticviol	8uv4cw	[5, 10]	I go from living happ	
19	745	anxiety	6f4swf	[0, 5]	I have a lot of self e	

Fig 2. Snapshot of Data set example [3]

EXISTING WORK ON DATASET

REVIEW OF PRE-PROCESSING TECHNIQUES

Elsbeth Turcan and Kathleen McKeown [3] worked on data analysis in two different categories. Like by domain, by the label.

By domain: By domain Authors applied to Examine each domain's vocabulary patterns using only our training data, excluding unlabeled data, to expand our analysis to the label level. First, we utilize the Linguistic Inquiry and Word Count (LIWC), a lexicon-based tool that provides scores for psychologically relevant categories such as melancholy or cognitive processes, as a proxy for topic predominance and expression variation.

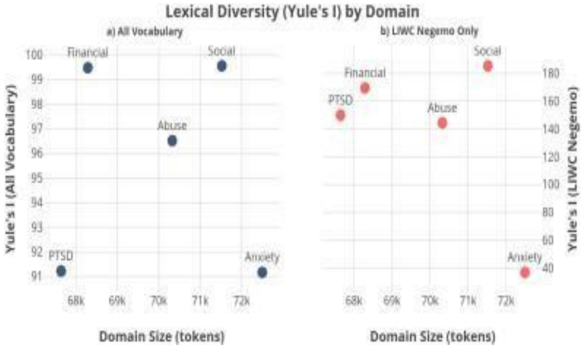


Fig3: Lexical Diversity by Domain [3]

By label: Authors conduct similar studies on data that has been classified as stressful or non-stressful by the majority of annotators. We validate certain typical findings in the mental health literature, such as the usage of more first- person pronouns in stressful data (possibly representing higher self-focus) and more social words in non-stressful data (maybe reflecting a better social support network). MICHAEL M. TADESSE, HONGFEI LIN, BO XU, AND

LIANG YANG [4] worked before proceeding to the feature selection and training stage, we use NLP tools to pre- process the dataset. To begin, we break the posts into Distinct tokens using tokenization. Next, we delete all URLs, punctuations, and stop words that, if left alone, could result in unpredictable results. Then we use stemming to reduce the words to their root form and group terms that are related together.

Here, the Author applied pre-processing technique removing common words, Words lemmatization technique was applied. [5]

Maryam Mohammed [7] Aldarwish, and Hafiz Farooq Ahmed collect a dataset from different generals like Face book, live Journals, and Twitter. Maryam Mohammed Al Darwish, Hafiz Farooq Ahmed has applied Rapid Miner was used to create the Predicting Depression Model. The model is made up of several processes that are used to test classifiers, SVM classifiers, and the Naive Bayes classifier. The first operator is Select Attributes, which determines which attributes of the training dataset should be retained and which should be removed. The second and third operators are the Nominal to Text, which changes the type of selected nominal attributes to text and maps all attribute values to corresponding string values; it is used in both the training and test datasets. The fourth and fifth processes are Process Documents, which are used in the training and test datasets to generate word vectors from string attributes and consist of four operators. Tokenize, Filter Stop words, Transform Cases, and Stem are the four Process Document operators. Tokenize operators convert a document`s text into a sequence of tokens.

Priyanka Arora.[8], Parul Arora collected data set from Twitter 3754 tweets, and they both applied pre-processing techniques like this

Case Folding: All the words in the dataset are converted to lowercase. The following data pretreatment procedures have been eliminated due to a lack of functionality. Unnecessary punctuations, Extra blank spaces, URLs, Hashtags

Replacing Emoticons: Emoticons are replaced with terms to represent different emotions, such as positive emoticons (EPOS), negative emoticons (ENEG), and neutral emoticons (ENEUT).

Removal of stopwords: Stop words are words that do not add meaningful content to the dataset (for example, pronouns, prepositions, and conjunctions). As a result, deleting them reduces the size of the objects in the training and testing set greatly.

Replacement of sentiment words: Words representing sentiments are replaced with keywords, such as POS for positive phrases and NEG for negative words.

Mrs. B. Ida Seraphim, Subroto Das, Apoorv Ranjan. [9] collect data from Kaggle and other government portals, then after they worked on.

Data cleaning: remove null values and unnecessary rows from the dataset after which Authors will do Min-max normalization or standardization.

REVIEW OF FEATURE EXTRACTION TECHNIQUES

The Authors applied three feature extraction techniques is applied Lexical features, Syntactic features, social media features, and Syntactic features. [3]

Lexical features: The full suite of 93 LIWC features; average, maximum, and lowest scores for pleasantness, activation, and imagery from the Dictionary of effect in Language (DAL); and sentiment assessed using the Pattern sentiment collection.

Syntactic features: Part-of-speech unigrams and bigrams, the Flesch-Kincaid Grade Level, and the Automated Readability Index.

Social media features: The post's UTC timestamp; the ratio of up votes to down votes on the post, where an up vote approximately correlates to a "like" reaction and a down vote to a "dislike" reaction. The posts net score (karma) (determined by Reddit, based on n up votes and down votes) here n mines no post.

In this paper Authors applied the feature extraction techniques is n-gram, liwc, LDA.[4]

N-gram: The features from the postings are examined using N-gram modeling. It's a characteristic for depression detection that's commonly utilized in text mining. N-gram feature also applied their method name is unigram & bigrams.

Example unigram & bigrams.

Table: 1 Example n-gram [4]

SL.No.	Type of n- gram	Generation-grams
1	Unigram	["I"," reside"," in", "And "]
2	Bigrams	["I reside", "in Anand"]

And NLP, for modeling authors, is used to (if-IDF) the full form Term frequency-inverse document frequency

Topic modeling: When the validation set is limited to 70 topics, the LDA model performs best. Only words that appear in at least 10 postings are considered for the topic selection. Every post is treated as a single document that must be tokenized and stemmed.

LIWC: To conduct Authors study, we selected 68 out of 95 features based on psycholinguistic measurements and converted each depressive and non-depressive post into numerical values. Authors can then calculate scores for three higher-level categories based on standard language dimensions, psychological processes, and personal concerns in this fashion. One of the most important aspects of language is the normal linguistic processes. The LIWC

Psycholinguistic Vocabulary Package is a collection of psycholinguistic vocabulary.

Feature type	Method	Number of selected Feature
N-grams	Unigrams Big grams	3000 2736
Linguistic Dimensions psychological process personal concern process	LIC	68
Top modeling	LDA	70

Table: 2. Different types of approaches to text encoding methods. [4]

Lizakonopelko [5] is applied as a feature extraction technique. "Term Frequency — Inverse Document Frequency" is abbreviated as TF-IDF. This is a method for calculating the number of words in a collection of documents. Authors usually assign each word a score to indicate its prominence in the document and corpus. In the fields of information retrieval and text mining, this method is commonly employed.

sohommajumder [6] is applied in feature extraction as a Bert tokenized. Priyanka [8] Arora, Parul Arora applied different kinds of feature extraction techniques Tokenization, Stemming, Gram features Sentiment Extractions and POS vector.

In this paper Mrs. B. Ida Seraphim, Subroto Das, Apoorv Ranjan [9] applied exploratory analysis for feature selection and the second technique applied Binary. Based on the suicide incidences per 100,000 people, Authors chose to undertake a binary classification of the suicide data, assigning high/low suicide risk groups. The 'Risk' column is added to the "full" data frame as an extra column.

Suicides<mean (Suicides) ->low risk --> class 0
Suicides>mean (Suicides) --> high risk --> class 1

3. A review of different classifiers used

In the paper applied Elsbeth Turcan, Kathleen McKeown[3] Support Vector Machines (SVMs), logistic regression; Nave Bayes, Perceptron, and decision trees are among the non-neural models we try first. The parameters are tweaked. Grid search and a 10-fold multiplier were used for these models. Cross-validation, and acquire results for a variety of situation Input and feature combinations. As well also applied CNN (Convolution Neural Network) & both a two-layer bidirectional Gated Recurrent Neural Network (GRNN) We try training embeddings with random initialization and initializing using our domain-specific Word2Vec embeddings, as well as concatenating the best feature set from our non-neural experiments onto the Representations after the recurrent and convolution/pooling layers. When applying BERT to any task, we apply it straight to it, fine-tuning the pre-trained BERT-base8 on our classification task for three epochs. The appendix contains the parameter settings for our various models.

In this paper MICHAEL M. TADESSE, HONGFEI LIN, BO XU, AND LIANG YANG[4] applied LR (logistic regression), Random forest, Adabost, SVM, mlp.]Here lizakonopelko [5] applied SVC(Support Vector Classifier) algorithms. Sohommajumder[6] has applied machine learning algorithms like LR, KNN, SVM, Xg boosts, Ada boosts, RF, Decision tree classifiers, GBM(Gradient Boosting Machine), Cart [7] In this paper, the Authors applied Machine learning algorithms Like Svm & Naïve babies with the parameter of Recall, Accuracy, and F1 score. In this paper priyanka arora ,parurl arora [8] applied SVM & Multinomial Naïve Bayes. In this survey paper. In this paper Author [9] applied Decision Tree, AdaBoost, XGBoost, and Neural Network is some of our Machine Learning techniques.

Result & Discussion

Models	Precision	Recall	F1 score
Majority baseline	0.51611	1.0000	0.6808
CNN + features*	0.60230	0.8455	0.7035
CNN*	0.5840	0.9322	0.7182
GRNN w/ attention + features*	0.6792	0.7859	0.7286
n-gram baseline*	0.7249	0.7642	0.7441
n-grams + features*	0.7474	0.7940	0.7700
LogReg w/ pretrained Word2Vec + features	0.7346	0.8103	0.7706
LogReg w/ fine- tuned BERT LM+ features*	0.7704	0.8184	0.7937
LogReg w/ domain Word2Vec + features*	0.7433	0.8320	0.7980
BERT-base*	0.7518	0.8699	0.8065

Table 3: Supervised Results [3]

Our best model achieves a 79.80 F1 score on our test set, comparable to the state-of-the-art pretrained BERT-base model. In this table, “features” always refers to our best-performing feature set (≥ 0.4 absolute Pearson’s r). Models marked with a * show an Excellent improvement

Over the majority baseline (approximate randomization test, $p < 0.01$). Authors applied different Models; the best model is a logistic regression classifier with Word2Vec embeddings. The model achieves s 79.80 F1 scores on our test set. Compared to the majority baseline, the n-gram baseline, and the pre-trained embedding model, there was a considerable improvement. Our logistic regression classifier outperforms BERT-base (approximate randomization test, $p > 0.5$) with the added benefits of better interpretability and less intensive training. Furthermore, as expected, domain-specific word embeddings trained on our unlabeled corpus (Word2Vec, BERT) outperform n-grams or Our logistic regression classifier outperforms BERT-base (approximate randomization test, $p > 0.5$) with the added benefits of better interpretability and less intensive training. Furthermore, as expected, domain-specific word embeddings trained on our unlabeled corpus (Word2Vec, BERT) outperform n-grams or pre-trained embeddings, highlighting the importance of domain knowledge in this task. Embeddings, highlighting the importance of domain knowledge in this task. Although our basic deep learning models do not outperform the majority baseline as well as our standard supervised models or BERT, they consistently exceed the majority baseline.

In this Paper Authors [4] applied different classification algorithms with different feature extraction techniques.

Bigrams acquire the highest accuracy, particularly with the SVM learning method, which achieves 80 percent accuracy and a 0.79 F1 score; followed by the LIWC feature an RF model (78 percent, 0.84), and LDA with LR text classifier (78 percent, 0.84). (77 percent, 0.83). Bigram works well when evaluating single features; however, it has severe drawbacks when considering combined features. In terms of predictive power, LIWC outperforms LDA as a single feature. The highest precision is achieved with the RF model. SVM (74 percent, 0.74) and MLP (70 percent, 0.72) were outperformed by SVM (78 percent, 0.84) and MLP (70 percent, 0.72). In contrast to past research, LDA outperforms LIWC in this study. To supplement the N- grams, we use LIWC and LDA. In our investigation, the LIWC+LDA+bigram and MLP neural network model had the best performance for detecting depression. It excels SVM (90 percent, 0.91), LR (89 percent, 0.89), and RF (85 percent, 0.85) and AdaBoost (85 percent, 0.85) with its 91 percent accuracy and 0.93 F1 score (79 percent, 0.81). Furthermore, it outperforms another feature set containing unigrams (LIWC, LDA, unigram) as the best technique, with RF as the best strategy (83 percent, 0.84). The predictive capacity for greater performance is concealed in good feature selections and many feature combinations, according to our findings Bigrams acquire the highest

accuracy, particularly with the SVM learning method, which Achieves 80 percent accuracy and a 0.79 F1 score; followed by the LIWC feature an RF model (78 percent, 0.84), and LDA with LR text classifier (78 percent, 0.84). (77 pcent, 0.83). Bigram works well when evaluating single features; however, it has severe drawbacks when considering combined features. In terms of predictive power, LIWC outperforms LDA as a single feature. The highest precision is achieved with the RF model. SVM (74 percent, 0.74) and MLP (70 percent, 0.72) were outperformed by SVM (78 percent, 0.84) and MLP (70 percent, 0.72). In contrast to past research, LDA outperforms LIWC in this study. To supplement the N-grams, we use LIWC and LDA. In our investigation, the LIWC+LDA+bigram and MLP neural network model had the best performance for detecting depression. It excels SVM (90 percent, 0.91), LR (89 percent, 0.89), RF (85 percent, 0.85) and AdaBoost (85 percent, 0.85) with its 91 percent accuracy and 0.93 F1 score (79 percent, 0.81). Furthermore, it outperforms another feature set containing unigrams (LIWC, LDA, unigram) as the best technique, with RF as the best strategy (83 percent, 0.84). The predictive capacity for greater performance is concealed in good feature selections and many feature combinations, according to our findings

Izakonopelko[5] got an accuracy of only 51% using the linear svc algorithm. sohommajumder [6] applied a supervised machine learning model.

Name algorithms	Accuracy (%)
LR	76%
Ken	65%
Cart	74%
Bayes	7%
Xg boots	76%
GBM	76%
RF	0.69
Ada boost	0.75

Table: 4 performances of supervised algorithms

Logistic regression, XGBoost, and SVM, GBM gave the highest accuracy. Bayes gives the lowest accuracy.

Maryam Mohammed Aldarwish, Hafiz Farooq Ahmed [7] here applied SVM algorithms and Naïve Bayes

Name algorithms	Accuracy	Precession	Recall
SVM	57%	67%	56%
Naïve Bayes.	63%	100%	58%

Table:-5 performance of supervised algorithms

After the analysis of the Review, the system Researchers proposed system gets excellent results compared to others.

Here show the table Priyanka Arora and Parulrora [8] applied SVR, Multinomial Naïve Bayes algorithm applied here Multinomial Naïve Bayes get 78% accuracy and SVR get 79%7.Performance applied Researchers[9] algorithms by Ad boost were the most accurate of the four machine learning models, with a score of 97.6 %, followed by xg- boost with a score of 96.35 percent. The decision tree had 91.67 %accuracy, and the map classifier had 90 % accuracy.

4. Conclusion

We all are leavening on the digital era. We all are habited to extensive use of social media; there so many mental stress-related health issues arise. Now if we talk about this survey Different techniques, different datasets, different authors, different objectives, and different results. There is a Pre-processing step is always performed to enhance the input of subsequent steps. Research on intermediate steps.[3] Here Researchers identified that their deep learning models are not working up to the mark in terms of performance. [4] Authors experiment reveals that the applied approaches perform quite well; yet, the absolute values of the metrics indicate that this is a difficult undertaking that merits additional investigation. Authors believe that this experiment will help to establish the foundation for new mechanisms to be used in various sectors of healthcare to evaluate depression and related characteristics. [7] Authors used Rapid Miner to test two classifiers (SVM and Nave Bayes Classifier) in a predating depression model. Using the same patients as before data that has been incorporated into the planned online application, as well as Using a training dataset, the data was manually categorized.2073 depressing posts and 2073 non-depressed posts were used. There For each of the three outcomes, performance has been calculated. The Naive Bayes results, the sentiment results, and the SVM result. [8] This research proposes a novel approach to identifying health tweets for sadness and anxiety from all mixed tweets, allowing us to determine health status in the real world. It's a new platform for patient interaction, with many of them taking part in decision-making for better healthcare treatment outcomes. For future work, a system that is more efficient and accurate than the current system can be built, and the project can be expanded to aid others. To assess the service quality of any healthcare facility the spread of diseases and the consequences of medicinal products study the experiences of others' tweets The system is capable of could be developed to include more than one language, for example, Urdu, Punjabi, and Hindi are three languages spoken in India. [9]According to Researchers Gradient boosted decision trees and neural networks, for example, routinely beat other algorithms and had the highest accuracy and precision. Chat bots had a poor accuracy rate, but AI is still in its infancy, and there is more Work to be done in the field of human emotion recognition. My point of view Related [3] [4] is Authors applied good feature extraction techniques and get accurate results in stress detection & analysis.

References

[1] Awasthi, P. (2020, December 11). 74% of Indians suffering from stress; 88% reported anxiety amid Covid-19: Study. The Hindu BusinessLine. https://www.thehindubusinessline.com/news/variety/74-of-indians-suffering-from-stress-88-reported-anxiety-amid-covid-19-study/article33306490.ece

[2] Mental health. (2019, December 19). WHO. https://www.who.int/health-topics/mental

[3] Turcan, E. and McKeown, K., 2019. Dreaddit: A Reddit dataset for stress analysis in social media. arXiv preprint arXiv:1911.00133.

[4] Tadesse, M.M., Lin, H., Xu, B. and Yang, L., 2019. Detection of depression-related posts in reddit social media forum. IEEE Access, 7, pp.44883-44893.

[5] L. (2021, July 2). mental health subreddits. Kaggle. https://www.kaggle.com/code/lizakonopelko/mental -health-subreddits

[6] S. (2021, July 3). BERT tokenizer with 9 models-NLP stress analysis.Kaggle. https://www.kaggle.com/code/sohommajumder21/b ert-tokenizer-with-9-models-nlp-stress-analysis

[7] Aldarwish, M.M. and Ahmad, H.F., 2017, March. Predicting depression levels using social media posts. In 2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS) (pp. 277- 280). IEEE.

[8] Arora, Priyanka, and ParulArora. "Mining Twitter data for depression detection." 2019 International Conference on Signal Processing and Communication (ICSC). IEEE, 2019.

[9] Mrs. B. Ida Seraphim , Subroto Das , Apoorv Ranjan, 2021, A Machine Learning Approach to Analyze and

[10] Predict Suicide Attempts, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH &TECHNOLOGY (IJERT) Volume 10, Issue 04 (April 2021)