Stress Detection System using Natural Language Processing and Machine Learning Techniques

Kirti Kumari[1,*], Sima Das[2]

[1]Indian Institute of Information Technology Ranchi, Namkum, Ranchi,Jharkhand, India.
[2]Bengal College of Engineering and Technology Bidhannagar, Durgapur, West Bengal, India.

Abstract

Stress is one of the important phase of mental state where we feel emotional or physical tension. If we do not find way to manage stress then it may leads to physical or mental health issues. The COVID-19 pandemic affected almost everyone into the stressed phase due to long duration of social distancing, lockdown, fear, negativity, etc. The people used more online activity as an alternative way of physical activity in the last three years. The Internet is an enormous canvas for people to post everything that they see in their daily lives. Social media is frequently used for analysis, judgment, inspiration, or emotion detection. The objective of this work is to analyze sentiment from social media data and detect stress of various class of people from their social networking platforms. The proposed work is having two different components: first extraction of information using natural language processing and another is stress detection using machine learning techniques. Proposed work has 4 main phases: collection of data from social media, auto summarization, text mining, and stress detection. The proposed model can predict stress or cognitive load of an online user. The current model has used various machine learning techniques among them Support Vector Machine is giving good results compared to others techniques, it has 90% accuracy, 90% recall, 94% precision and F1 score is near about 92%. The current model will have a positive impact on society for the early detection of stress.

Keywords

Stress Detection, Natural Language Processing, Machine Learning, ICON 2022, WNLPe-Health 2022

1. Introduction

Stress is a normal feeling which can experience by anyone irrespective age group, gender or community. Most important thing is that how we are reacting with stress that will matter a lot [1]. The stress can be negative or positive. Especially COVID-19 pandemic affected almost everyone into stressed phase due to several restrictions like very long duration of social distancing, lockdown, etc. Due to those restriction people spend a lot of time on social media during pandemic period [2, 3]. These media is applicable almost in every discipline, to name some; we've, balloting mechanism for splendor festival, political campaigns, product research and promoting via advertisements. There is need for analyzing and modeling of such networks. Technology place is growing at a completely fast tempo leading to formation of new

19th International Conference on Natural Language Processing (ICON 2022): WNLPe-Health 2022, December 15–18, 2022, IIIT Delhi, India

kirti@iiitranchi.ac.in (K. Kumari); simadascse@gmail.com (S. Das)
https://iiitranchi.ac.in/ (K. Kumari)
0000-0003-3714-7607 (K. Kumari)

sophisticated gear from textual contents. Data mining techniques are required for his or her functionality of managing the three dominant residences with social network facts particularly; length, noise and dynamism. This huge quantity of information of social community requires automation for dispensation of data, reading it within a stipulated time. Interestingly, statistics mining techniques are designed to handle the voluminous statistics sets to mine extensive styles from statistics; social network sites provide these massive records sets due to their utilization and consequently are best applicants to mine information using the statistics mining equipment. Therefore, we will infer that the statistics mining or to be specific web mining provides the vital intelligence to the social network to create and engage in a more humanly and person friendly manner. Automatic stress detection is emerging as an appealing research area due to the growing demands for communication between intelligent systems and humans.

Researchers have designed a variety of techniques for analyzing physiological data obtained from sensors attached to human bodies in order to identify stress and categorize emotions [4, 5]. But there is very few works which focused on analyzing social media posts for analyzing the stress of online users. Here, we have tried to analyze the online user behavior styles on big-scale social networks or even use such social records for further studies. With the help of Machine Learning algorithms our system will provide accurate and reliable outcomes towards early detection of stress faced by the online users.

In this study, we propose a system for detecting psychological stress based on online textual comments. The idea of this work is to come up with natural language processing with machine learning that not only detects stress but also analyses the topic of debate in a specific social media text for further mental illness. Along with sentiment analysis, this mechanism will be useful in examining and segregating the consumer's opinions on other special subjects. After wearing out in-depth studies on pertinent datasets we are able to gain vital understandings of various correlations between social interactions and the anxiety or pressure of the online user.

Contribution of the paper is having following key points:

- Natural language processing techniques used to extract information from the comment.
- Analyze and segregate the person's reviews on specific topics.
- Stress detection using different types of machine learning methods and also choose an appropriate method for detection stress with the good accuracy.

The rest of the paper is organized into five different sections. Section 2 for literature survey on similar task and similar techniques, Section 3 discusses the proposed work and Section 4 for results and discussion. Finally, we concluded the paper with discussing about future scope in the Section 5.

2. Literature Survey

Social media has grown significantly during the last ten years, both positively and negatively. People may now contact each other directly across any cultural or economic divide because of the rapid growth of networking through social media and the internet. Social media has numerous advantages [6, 7], but it also has drawbacks that have a bad impact on society.

The active areas of research community are detection of hate, offensive, aggressive and stress related comments. Stress detection is not a new area of research, there are numerous research done in the field [8, 9, 4, 5, 10, 11, 7]. Hate speech and Online Aggression are issues that have emerged during the past few years [12, 13, 14, 15]. The use of derogatory and abusive remarks on social media is essentially considered hate speech. It could be referring to a single person or a certain group of people who share the same interests. In this essay, we have outlined our strategy for combating hate speech and significantly reducing it. People express their wrath and rage on social media immediately, which is hurtful to other people's feelings. It would be extremely detrimental to them and harm their caste, creed, religion, and race. Despite not necessarily intending to offend anyone, some comments could be considered hate speech because of the profanity they include. To eradicate hate speech, authors [16] have dig deeply into natural language processing and used a variety of machine learning models to determine which one to deploy based on its accuracy.

Our daily lives involve the linkage and retrieval of information. Databases are the most widely used source of information online. Data volume is increasing quickly, and database technology is advancing and having a significant impact. Virtually all web apps use databases to store and retrieve data. The objective is to offer a more user-friendly mechanism for creating database queries and delivering results. It is feasible to communicate with a large segment of the people using social media. They [17] increased public access to the Atmospheric Radiation Measurement Data Center (ADC) data using this medium and Natural Language Processing (NLP).

The main contributor to disability and a major factor in suicide is depression. In recent years, social media has become one of the most popular ways to share information online. Most people communicate their thoughts, philosophies, and personal experiences on the Internet. The language people use on social media shows that depression has an impact on how they communicate. Because of the enormous rise in mental health awareness, it is now very important to recognize mental disease. In the study [18], authors looked at Twitter users' tweets for any signs of depression. They employed text mining and natural language processing strategies to do this. With the CNN and LSTM classification method, they were able to achieve an accuracy of 92%. Additionally, they evaluated their model against classifiers for logistic regression and TF-IDF. Here, we have summarised some of the important recent works [19, 20, 21, 22, 23, 24, 9, 25, 26] done in the area of stress detection with their consideration of contents, techniques and task with comparison of our current system in the Table 1.

It has been extensively studied how to identify stress using physiological signs. In almost all previous methods, physiological signals from sensors for the electrocardiogram, electrodermal activity, and electromyography were examined. These methods analysed physiological signals using conventional machine learning algorithms to detect stress and categorise emotions. In the current work, we have utilize the social media posts for the detection of stress, discussed in subsequent section.

Table 1
Comparison table for recently published paper with proposed system based on Stress detection

Source	Measurement	Techniques	Task
Subhani et al. [19]	T test, Distance	Logistic regression, Support vector machine, Naïve bayes	Mental stress detection
Elzeiny and Qaraqe [20]	Electroencephalogram, Hibert-Huanf Transform	K-Nearest Neighbor	Workplace stress detection
Papini et al. [21]	Medical and demographic features	Logistic regression	Posttraumatic stress detection
Jadhav et al. [22]	Textual data, facial expression	Bidirectional Long Short-Term Memory	Text based stress detection from social media.
Dubey et al. [23]	Assisted reproductive technology	Support vector machine	Human spermatozoa detection under oxidative stress
Jebelli et al. [24]	Electroencephalogram	Online Multi-Task Learning (OMTL) algorithms	Stress recognition framework
Das et al. [9]	Electroencephalogram	Backpropagation Neural Network	Cognitive load detection
Zhang et al. [25]	Magnetoencephalography, Electoencephalogram	Support vector machine	Posttraumatic stress detection
Yousefi et al. [26]	Pupildiameter, electrodermal activity	Linear discriminant analysis	Stress detection using eye tracking dataset
Our proposed system	Textual data	Natural language processing with Support vector machine.	Text based stress detection from Twitter.

3. Methodology

In this section, we have discussed the dataset collection and annotation in the first subsection. Then, in second subsection, we discussed the proposed approach.

3.1. Dataset Collection and Labelling

We have collected our dataset by using Tweepy API from the Twitter, which is most frequently used social media by the online user. For the initial filtration, we have applied words based identification of Tweets and selected some of the users which are more vulnerable for the stress. Then, we collected their tweets from last three months and taken for our case study. We have found 58 users and their total 2978 tweets are selected for our study. Here, we have shown the simple examples of stressed people reaction, which can be seen in Figure 1. Similarly, we have shown in Figure 2, some of the important examples of stress related Tweets from our dataset. We can observe from the Figures 1 and 2, physical stress is more explicitly can be seen but stress of online users are more implicit, which making the task challenging.

Then, we have manually labelled the tweets into low stressed verses high stressed by the help of five undergraduate and two postgraduate students. All seven students labelled the tweets

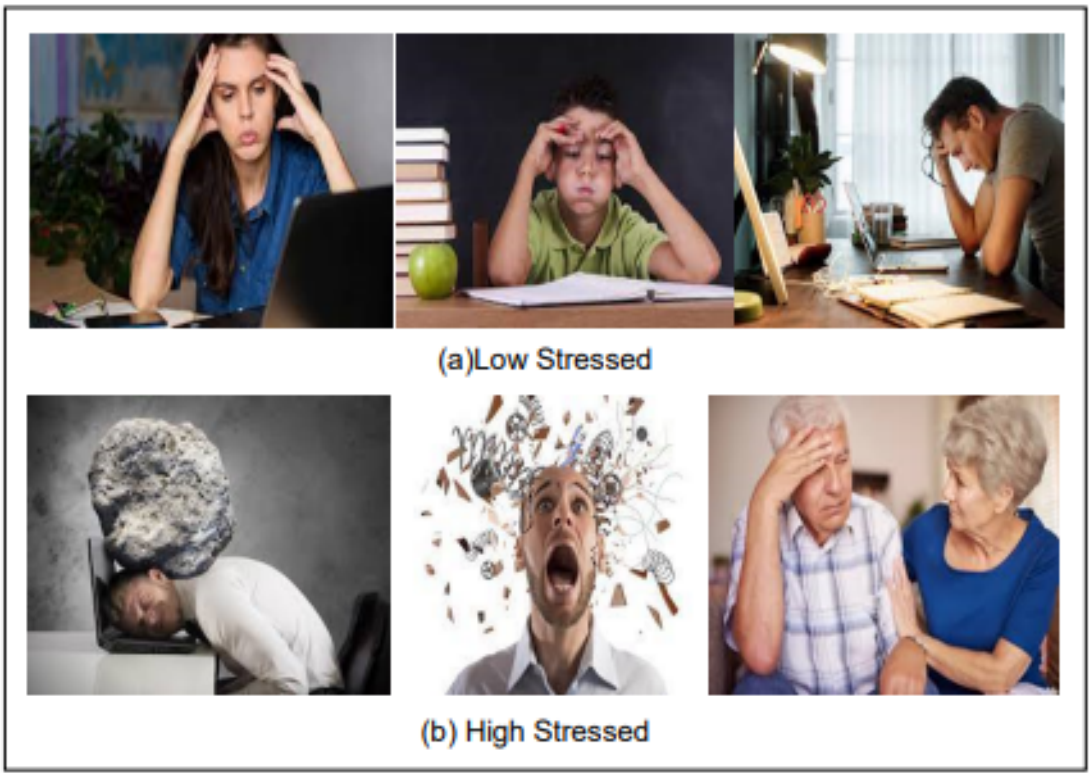

Figure 1: Examples of physical images of Stress

independently. Finally, we have considered the ultimate label with the help majority voting strategy. The detail distribution of different stress class data shown in Figure 3.

3.2. Proposed Approach

In this section, we have presented our approach towards stress detection through natural language processing and machine learning techniques.

Unstructured data, which makes up almost 80% of all data in the world, is analyzed and processed in one of the most crucial ways possible through text mining. Most organizations and institutions now collect and store enormous volumes of data in data warehouses and cloud platforms, and as fresh data floods in from various sources, this data continues to expand dramatically by the minute. Overall, our proposed work is shown in Figure 4, which is designed with natural language processing with machine learning techniques used to detect stress over social media extracted information. The proposed system has 4 main phases, first one for dataset collection from social media, 2nd phase designed for auto summarization of all the post collected from one specific user shown in Figure 5, 3rd phase designed for text mining which is shown in Figure 6 and last phase is designed to detect stress that phase shown in Figure 7.

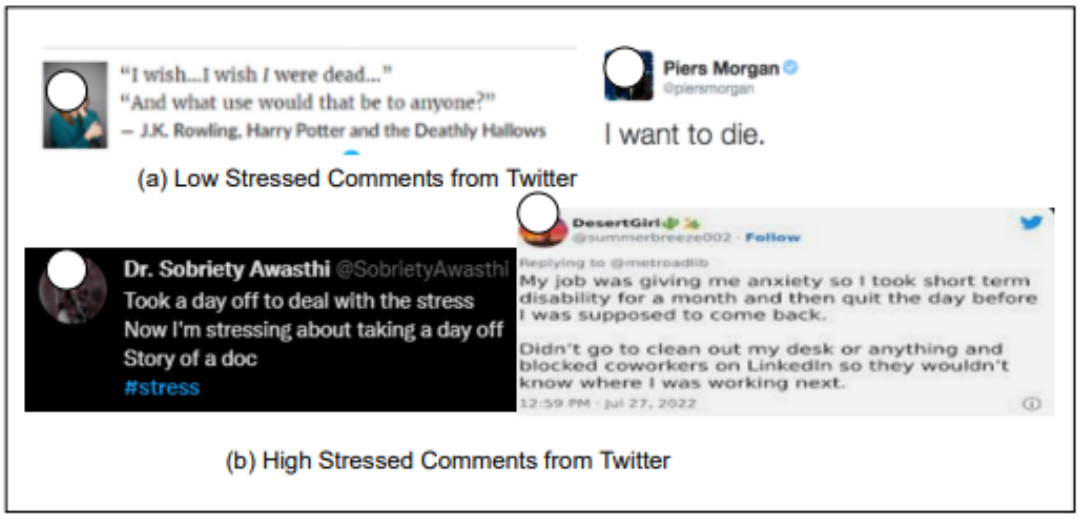

Figure 2: Examples of Stress related Tweets

Figure 3: Distribution of stress data

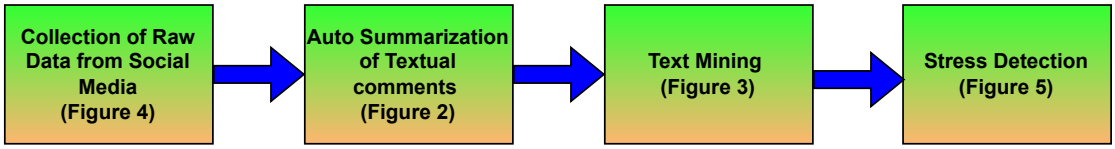

Figure 4: Overview of our proposed work

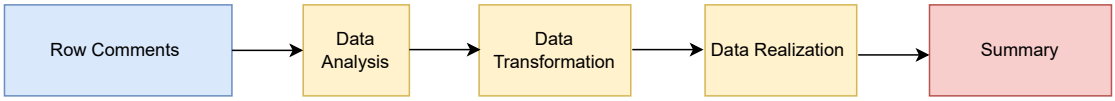

Figure 5: Auto Summarization of Comments

3.3. Automatic Summarization

Automatic summarization as shown in Figure 5 is the act of computationally condensing a set of data to produce a subset (a summary) that captures the key ideas or information within the original text. Here, we have taken all the comments from the same online user and summarized the comments. The main idea of summarising the comments is that every comment is not equal important for the stress detection. The approach presented uses K-Means clustering to create extracts from the parent text. The number of clusters depend upon the size of the input text. In broader view though, too less number of clusters are unsuitable as they may change the meaning of the parent text entirely. Similarly large number of clusters would mean that the size of the extracted text is large which contradicts the purpose of summarization.

3.4. Text Mining

For extracting information from the textual comments, we have utilized the text mining techniques shown in Figure 6. At first we have tokenized the corpus. Then, we used stemming and lemmatization methods for better understanding of root word or token as normalization technique. Further, we have used entity recognition for the important terms related to stress and taken as a separate feature. The most common data pre-processing activity is named entity recognition (NER). It entails locating important information in the text and classifying it into a number of predetermined categories. A constant subject of discussion or reference in a book is referred to as an entity. These extracted information given to the machine learning classifier for differentiating the low stress verses high stress comments.

Natural language processing (NLP) is an artificial intelligence technique that turns available (unstructured) text found in documents and databases into normalized, structured data that can be used for analysis or as input for machine learning algorithms. Figure 6 shows the proposed model for text mining. Unstructured data, which makes almost 80% of all data in the world, is analyzed and processed in one of the most crucial ways possible through text mining.

Finally, we have come up with extracted data and with their two labels: Low Stressed and High Stressed. With these information we have experimented with different classifiers. Here, we have taken 80% samples for the training and rest 20% for the testing. We have experimented with two different vectoriser techniques: Count-Vectoriser and Term Frequency-Inverse Document Frequency (TF-IDF) vectoriser for making suitable input for our classifier. Latter, we found that TF-IDF is performed well so we left the Count-Vector for latter stage. We have used NLTK library[1] and Scikit-learn[2] package for the implementation.

[1]https://www.nltk.org/
[2]https://scikit-learn.org/

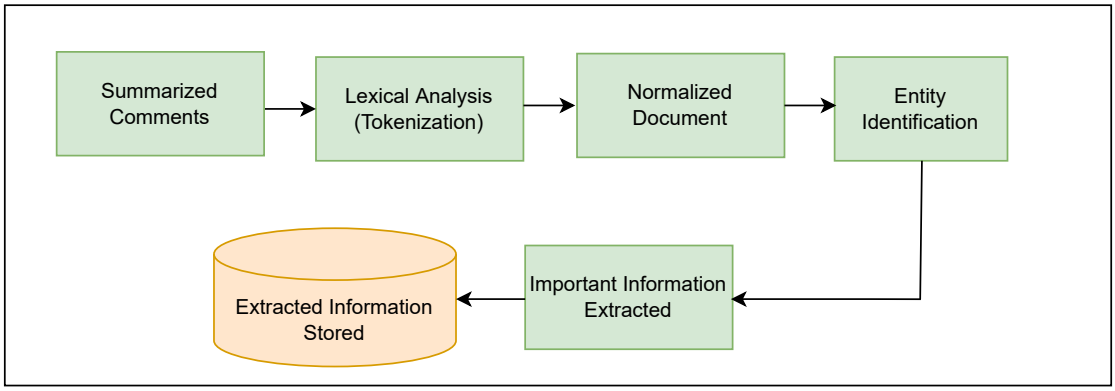

Figure 6: Proposed model for Text Mining

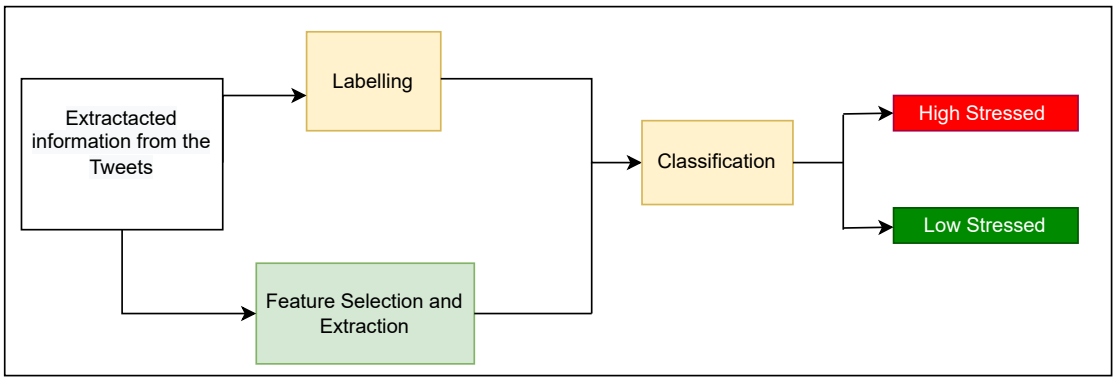

Figure 7: Proposed framework for Stress Detection

4. Result and Discussion

In this section, we presented our findings and analysis of experimentation with highlighting the limitations. In order to avoid a person from experiencing numerous stress-related health issues, this research provides automated system for stress detection on persons utilizing social media posts gathered through Twitter API and applying Natural Language Processing and using various machine learning algorithms. We have also tried basic prepossessing like removing stop-words, removing URLs, lowercasing the uppercase, etc. Specially, we have experimented with five machine learning classifiers: Support Vector Machine (SVM), Logistic Regression, Naive Bayes, Decision Tree and Random Forest classifiers. We have used Term Frequency-Inverse Document Frequency (TF-IDF) vectoriser for making input to the classifier. For the evaluation of our system, we have chosen the Accuracy, Precision, Recall and F1-Score as performance matrices. The Table 2 shows the performance of different classifiers and found that Support Vector Machine and Random Forest classifiers are performing better than other classifiers in

our case of experimentation.

Table 2
Performance of Proposed Framework

Classifier	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	**0.90**	**0.94**	**0.90**	**0.92**
Logistic Regression	0.89	0.92	0.88	0.89
Naïve Bayes	0.82	0.86	0.83	0.84
Decision Tree	0.78	0.81	0.76	0.79
Random Forest	**0.91**	**0.93**	**0.87**	**0.90**

We have not tried the deep learning models due to the insufficiency of samples. We have also not-considered the the tweets which has other than English word and having different modalities of data while choosing the tweets for the dataset.

5. Conclusion and Future Work

Most people have to deal with stress on a regular basis. However, long-term stress, or a high level of stress, will jeopardize our safety and disrupt our usual lives. Identifying mental stress proactively can help to prevent many health problems related with stress. In order to identify stress more effectively, this study attempts to offer a method for analyzing the mental stage during stress, based on social media posts made under stressful circumstances. With the help of the proposed method it is possible to identify people that have anxiety and depressive illnesses by utilizing prediction models to identify user language on social media, which has the potential to supplement conventional screening. Predictive models based on machine learning technique may provide the possibility to diagnose symptoms sooner, perhaps before psycho-social effects become serious.

In this work, we have experimented with very small volume of data but can be extended for large volume data for better understanding of stress of online users. Another future scope of our work is that it can be experimented with multi-lingual posts and multi-modal posts like image, Meme, audio and video; which is very common in India and also upgraded with features of social media.

References

[1] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, M. Tsiknakis, Review on psychological stress detection using biosignals, IEEE Transactions on Affective Computing 13 (2022) 440–460. doi:10.1109/TAFFC.2019.2927337.

[2] U. Naseem, I. Razzak, M. Khushi, P. W. Eklund, J. Kim, Covidsenti: A large-scale benchmark twitter data set for covid-19 sentiment analysis, IEEE Transactions on Computational Social Systems 8 (2021) 1003–1015.

[3] P. Gupta, S. Kumar, R. R. Suman, V. Kumar, Sentiment analysis of lockdown in india during covid-19: A case study on twitter, IEEE Transactions on Computational Social Systems 8 (2020) 992–1002.

[4] S. Greene, H. Thapliyal, A. Caban-Holt, A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health, IEEE Consumer Electronics Magazine 5 (2016) 44–56.
[5] Y. S. Can, B. Arnrich, C. Ersoy, Stress detection in daily life scenarios using smart phones and wearable sensors: A survey, Journal of biomedical informatics 92 (2019) 103139.
[6] S. Dosani, C. Harding, S. Wilson, Online groups and patient forums, Current Psychiatry Reports 16 (2014) 1–6.
[7] K. Kumari, S. Srivastav, R. R. Suman, Bias, threat and aggression identification using machine learning techniques on multilingual comments, in: Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022), Association for Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 30–36. URL: https://aclanthology.org/2022.trac-1.4.
[8] F.-T. Sun, C. Kuo, H.-T. Cheng, S. Buthpitiya, P. Collins, M. Griss, Activity-aware mental stress detection using physiological sensors, in: International conference on Mobile computing, applications, and services, Springer, 2010, pp. 282–301.
[9] S. Das, L. Ghosh, S. Saha, Analyzing gaming effects on cognitive load using artificial intelligent tools, in: 2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), IEEE, 2020, pp. 1–6.
[10] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ hasoc 2020: Bert models for hate speech identification in indo-european languages., in: FIRE (Working Notes), 2020, pp. 319–324.
[11] K. Kumari, J. P. Singh, Ai_ml_nit_patna@ trac-2: deep learning approach for multi-lingual aggression identification, in: Proceedings of the second workshop on trolling, aggression and cyberbullying, 2020, pp. 113–119.
[12] K. Kumari, J. P. Singh, AI_ML_NIT_Patna @ TRAC - 2: Deep learning approach for multi-lingual aggression identification, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 113–119. URL: https://aclanthology.org/2020.trac-1.18.
[13] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, Evaluating aggression identification in social media, in: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 1–5. URL: https://aclanthology.org/2020.trac-1.1.
[14] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[15] K. Kumari, J. P. Singh, Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content., FIRE (working notes) 2517 (2019) 328–335.
[16] R. Devarakonda, M. Giansiracusa, J. Kumar, H. Shanafield, Social media based npl system to find and retrieve arm data: Concept paper, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 4736–4737.
[17] S. Chiramel, D. Logofătu, G. Goldenthal, Detection of social media platform insults using natural language processing and comparative study of machine learning algorithms, in: 2020 24th International Conference on System Theory, Control and Computing (ICSTCC), IEEE, 2020, pp. 98–101.
[18] M. Häberle, M. Werner, X. X. Zhu, Building type classification from social media texts via geo-spatial textmining, in: IGARSS 2019-2019 IEEE International Geoscience and Remote

Sensing Symposium, IEEE, 2019, pp. 10047–10050.
[19] A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel, A. S. Malik, Machine learning framework for the detection of mental stress at multiple levels, IEEE Access 5 (2017) 13545–13556.
[20] S. Elzeiny, M. Qaraqe, Blueprint to workplace stress detection approaches, in: 2018 International Conference on Computer and Applications (ICCA), IEEE, 2018, pp. 407–412.
[21] S. Papini, D. Pisner, J. Shumake, M. B. Powers, C. G. Beevers, E. E. Rainey, J. A. Smits, A. M. Warren, Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization, Journal of anxiety disorders 60 (2018) 35–42.
[22] S. Jadhav, A. Machale, P. Mharnur, P. Munot, S. Math, Text based stress detection techniques analysis using social media, in: 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), IEEE, 2019, pp. 1–5.
[23] V. Dubey, D. Popova, A. Ahmad, G. Acharya, P. Basnet, D. S. Mehta, B. S. Ahluwalia, Partially spatially coherent digital holographic microscopy and machine learning for quantitative analysis of human spermatozoa under oxidative stress condition, Scientific reports 9 (2019) 1–10.
[24] H. Jebelli, M. M. Khalili, S. Lee, A continuously updated, computationally efficient stress recognition framework using electroencephalogram (eeg) by applying online multitask learning algorithms (omtl), IEEE journal of biomedical and health informatics 23 (2018) 1928–1939.
[25] J. Zhang, J. D. Richardson, B. T. Dunkley, Classifying post-traumatic stress disorder using the magnetoencephalographic connectome and machine learning, Scientific reports 10 (2020) 1–10.
[26] M. S. Yousefi, F. Reisi, M. R. Daliri, V. Shalchyan, Stress detection using eye tracking data: An evaluation of full parameters, IEEE Access (2022).