IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134121

Social Media-Based Hate Speech And Stress Identification Through Machine Learning And Natural Language Processing (NLP)

Mrs. Sharon D’Souza[1], Ashwin Shetty[2], Jeevan M[3], Nishal SP Karkera[4], Rahul D Shetty[5]

Assistant Professor, Computer Science and Engineering, AJIET, Mangalore, India[1]

Student, Computer Science and Engineering, AJIET, Mangalore, India[2-5]

Abstract: The proliferation of hate speech on social media has become a pressing societal concern, prompting the need for effective identification and mitigation strategies. This abstract outlines a novel approach utilizing machine learning (ML) and natural language processing (NLP) techniques to detect hate speech and assess its impact on inducing stress among users. The study focuses on the development of an ML-based model trained on a diverse dataset of social media content to accurately identify hate speech. Leveraging NLP, the model aims to comprehend linguistic nuances, context, and sentiment within textual data, enabling it to distinguish between normal discourse and potentially harmful language. Furthermore, the research extends beyond mere identification, aiming to gauge the psychological impact of hate speech by analyzing its correlation with stress levels among social media users. By employing sentiment analysis and stress identification algorithms, the study aims to quantify the emotional toll experienced by individuals exposed to such content. The abstract emphasizes the interdisciplinary nature of the research, bridging the gap between computer science, linguistics, and psychology. The proposed methodology holds promise in aiding social media platforms, policymakers, and mental health professionals in devising targeted interventions to combat hate speech and mitigate its adverse effects on users' well being. Through this holistic approach, this study endeavors to contribute to the development of proactive strategies for early detection, intervention, and support mechanisms, fostering a safer and healthier online environment for all users.

Keywords: Stressfull comments, hate speech, personal assaults, healthier online environment.

I. INTRODUCTION

Over the last decade social media has acquired a lot of attraction both positively and negatively way with the fast growth of social networking. People can communicate with one other via many social media platforms. In these the digital age, social media platforms have become vibrant hubs for communication, fostering connectivity and information sharing. However, alongside these benefits, they have also become breeding grounds for hate speech and stress-inducing content. The pervasive nature of online interactions has highlighted the urgent need to develop robust mechanisms for identifying and mitigating harmful discourse.

Addressing this challenge requires innovative solutions that leverage machine learning to detect hate speech and stress indicators embedded within the vast expanse of social media content. This research endeavors to create a sophisticated machine learning framework capable of discerning hate speech and stress markers within the complex fabric of online conversations. By employing cutting-edge algorithms in Machine learning, natural language processing (NLP) and sentiment analysis, this system aims to detect subtle linguistic cues, contextual nuances, and patterns indicative of hate speech or emotional distress.

Through the analysis of user-generated content across various social media platforms, this technology seeks to categorize and flag harmful discourse, enabling swift interventions to maintain a healthier online environment. This project addresses the pervasive challenges of hate speech and stress on social media platforms through an innovative machine learning solution.

Leveraging a multi-modal approach that analyzes text, and images, our system dynamically adapts to changing sentiment landscapes. It goes beyond traditional sentiment analysis by prioritizing contextual awareness, considering broader conversations, user interactions, and historical context. Transparency is emphasized through explain ability in model predictions, fostering user trust

II. PROBLEM STATEMENT

Identifying hate speech and stress on social media through machine learning (ML) and natural language processing (NLP) involves tackling a multifaceted challenge that intersects technological, ethical, and social considerations. Ethical considerations play a pivotal role in this domain. Handling sensitive content like hate speech or distressing language demands a delicate balance between ensuring user safety and privacy while also mitigating potential biases within the algorithms. Ensuring fairness, transparency, and accountability in the models' decision-making processes is essential. Developing robust ML models involves creating algorithms that can learn patterns and features indicative of hate speech or stress.

This requires an extensive, diverse, and carefully annotated dataset for training. These datasets should encompass a wide range of demographics, languages, and social contexts to ensure the models are sensitive to various cultural nuances and aren't biased towards specific groups. One significant challenge is adapting to the constantly evolving nature of language and online behavior. New words, phrases, or expressions frequently emerge, necessitating continuous model updates and retraining to maintain accuracy and relevance. In conclusion, the identification of hate speech and stress on social media via ML and NLP is a complex and interdisciplinary challenge. Developing accurate, fair, and culturally sensitive models that can navigate the subtleties of language and user intent is pivotal to fostering a safer and more supportive online environment.

III. OBJECTIVE

1.The objective is to develop a robust machine learning system that effectively discerns hate speech and stress markers in social media content.

2. By leveraging algorithms capable of analyzing linguistic patterns and contextual cues, this system aims to accurately detect and categorize harmful or distress-inducing language.

3. Finding the stressful posts on social media platform which may causes psychological instability of people’s.

4. Data-Driven Decision Making Utilize AI/ML algorithms to analyze feedback data, extract meaningful patterns, and provide actionable insights. This objective aims to empower decision-makers with data-driven information for strategic human resource planning.

5. Promote Continuous Improvement: Establish a mechanism for continuous improvement by encouraging ongoing feedback loops. safer online environment by promptly identifying and addressing hate speech while also providing timely support for individuals facing stress or mental health challenges exacerbated by social media interactions.

IV. REQUIREMENT SPECIFICATION

Hardware Requirements

1. Processor: Intel(R) Core i3 & above Versions.
2. System Type: 64-bit operating system, x64-basedprocessor
3. Installed Ram: 8 GB
4. Network Infrastructure: High-speed and reliable network connections to ensure seamless communication between server components and responsiveness for end-users

Software Requirements

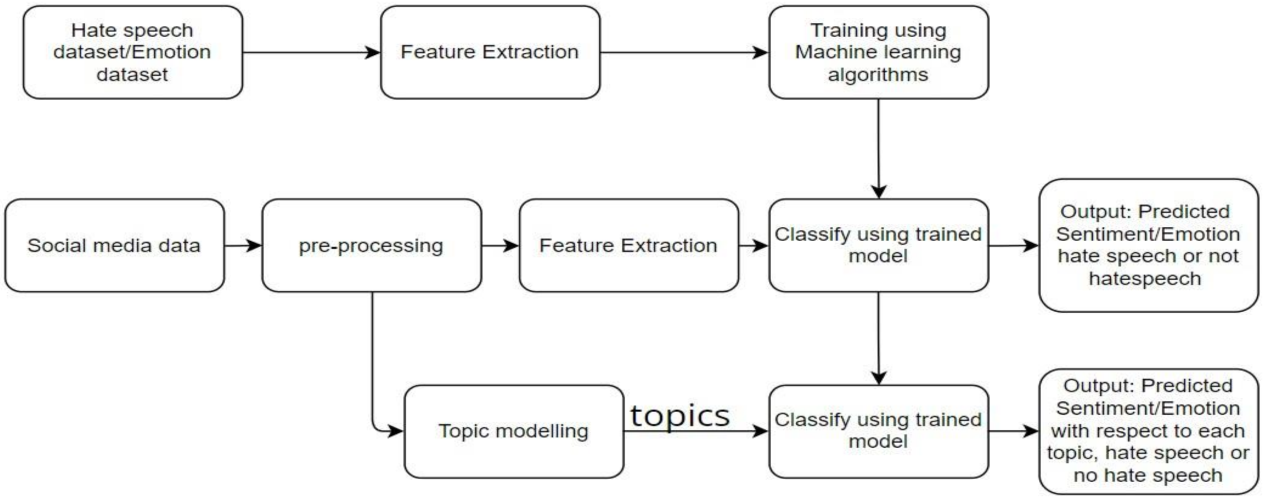
1. Platform and Hosting: Hosting for scalability, reliability, and security. Compatibility across various platforms (web, mobile, etc.).
2. Programming Languages and Frameworks: Python for AI/ML algorithms and backend development. TensorFlow or PyTorch for machine learning models.
3. Operating System: Choose a stable and secure operating system that aligns with government IT policies. Common choices include Linux distributions (e.g., CentOS, Ubuntu) or Windows Server.
4. Front-End Technologies: Implement front-end technologies (HTML, CSS, JavaScript) to create an intuitive and user-friendly interface. Consider using a front end framework.

DOI: 10.17148/IJARCCE.2024.134121

V. SYSTEM DESIGN

A. METHODOLOGY

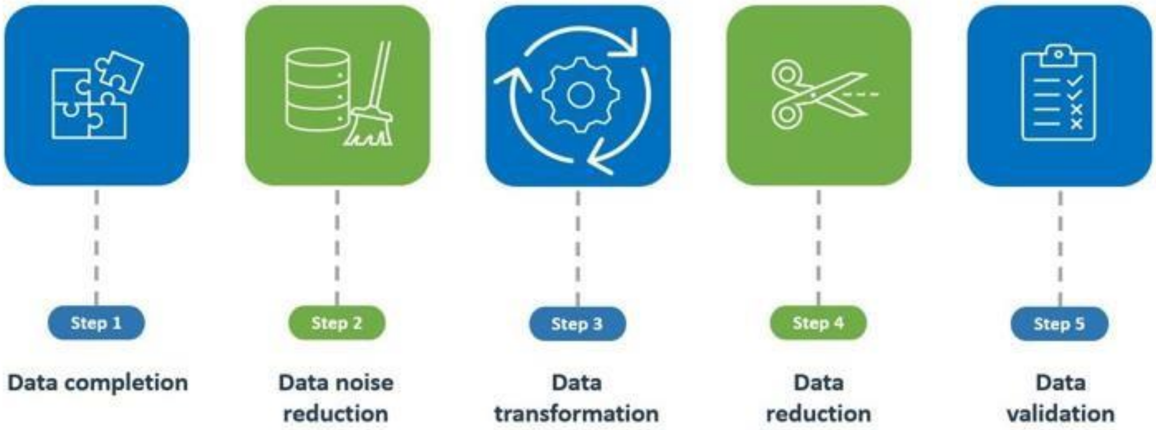
The Figure shows the steps that are followed in the methodology of this project.


Hate speech dataset/Emotion dataset
Feature Extraction
Training using Machine learning algorithms
Social media data
pre-processing
Feature Extraction
Classify using trained model
Output: Predicted Sentiment/Emotion hate speech or not hatespeech
Topic modelling
topics
Classify using trained model
Output: Predicted Sentiment/Emotion with respect to each topic, hate speech or no hate speech

1. Collect a hate speech and stress dataset: You will need a dataset of labelled examples of hate speech and non-hate speech. There are many publicly available datasets that you can use for this purpose, such as the Hate Speech and Offensive Language, also stress and no stress dataset or the Twitter Hate Speech dataset and stress dataset.

2. Pre-processing the data: Pre-processing involves cleaning and transforming the raw text data into a format that the machine learning algorithm can use. Some common preprocessing steps include tokenization, stop word removal, and stemming. Raw, real-world data in the form of text, images, video, etc., is messy.

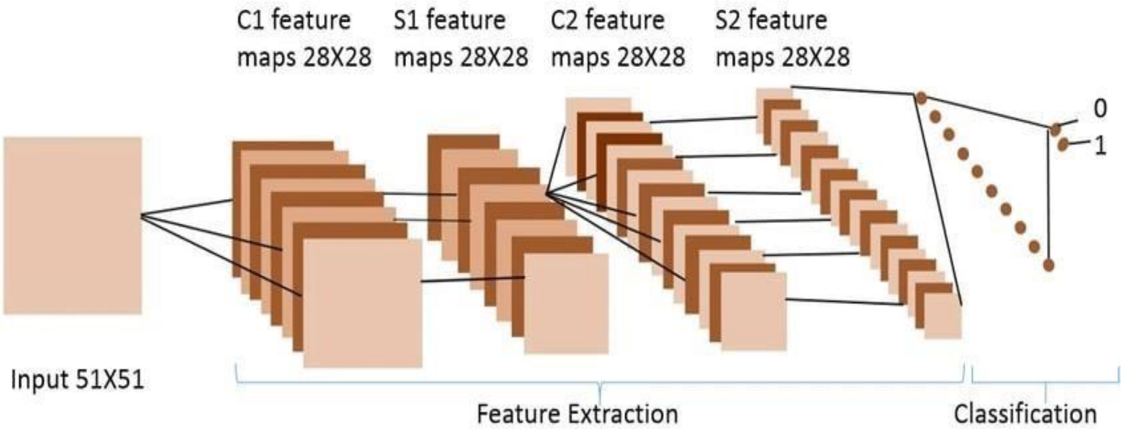
This Figure shows the steps that are followed in preprocessing steps of this project



Data preprocessing

3. Feature extraction: This step involves extracting relevant features from the preprocessed text. You can use techniques such as a bag of words, TF-IDF, or word embedding to create features that can be used by the machine learning algorithm. This Figure describes the feature extraction process used in the methodology of this project

DOI: 10.17148/IJARCCE.2024.134121



Future extraction

4. Train the model: Divide your dataset into training and validation sets. Use the training set to train your machine learning model. SVM and Naive Bayes are popular choices for hate speech and stress because they are relatively easy to implement and can work well with high-dimensional sparse feature vectors.

5. Evaluate the model: Use the validation set to evaluate the performance of your model. Common evaluation metrics include precision, recall, F1 score, and accuracy.

6. Deploy the model: Once you have trained and evaluated your model, you can deploy it to classify new text as hate speech or non-hate speech and stress or no stress.

B. ALGORITHM USED

The algorithms used in this project:
- MultinominalNB

MultinominalNB :
The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an comments or a social m, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

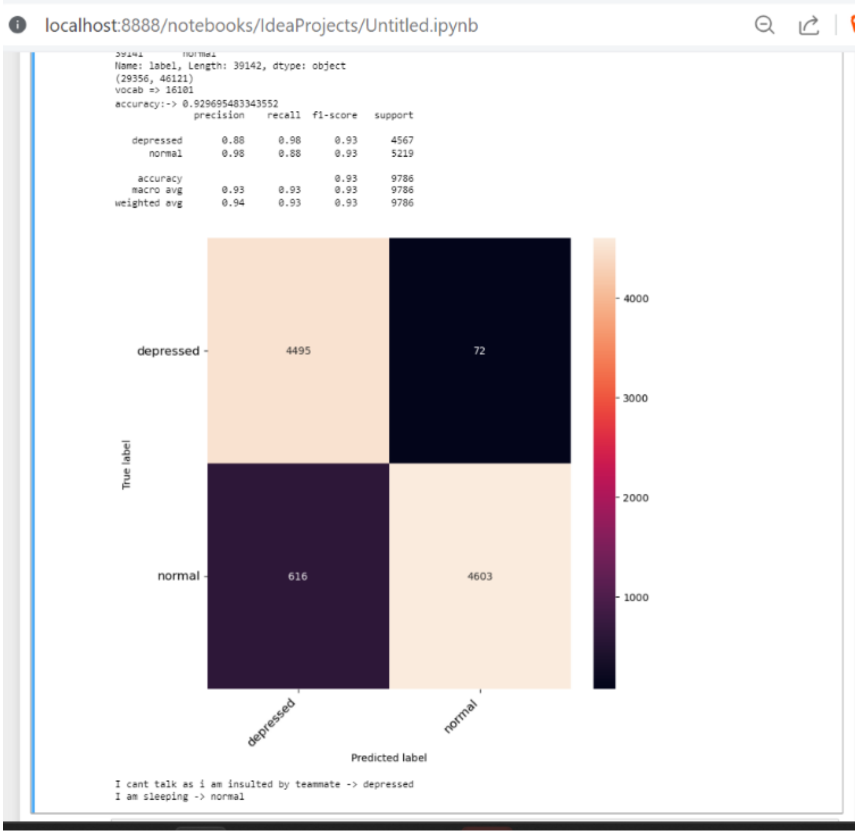
Naive Bayes is a probabilistic algorithm family based on Bayes' Theorem. It's "naive" because it presupposes feature independence, which means that the presence of one feature does not affect the presence of another (which may not be true in practice).

Multinomial Naive Bayes is a probabilistic classifier to calculate the probability distribution of text data, which makes it well-suited for data with features that represent discrete frequencies or counts of events in various natural language processing (NLP) tasks.

VI. OUTPUT

Below figure shows the heatmap of this project

DOI: 10.17148/IJARCCE.2024.134121



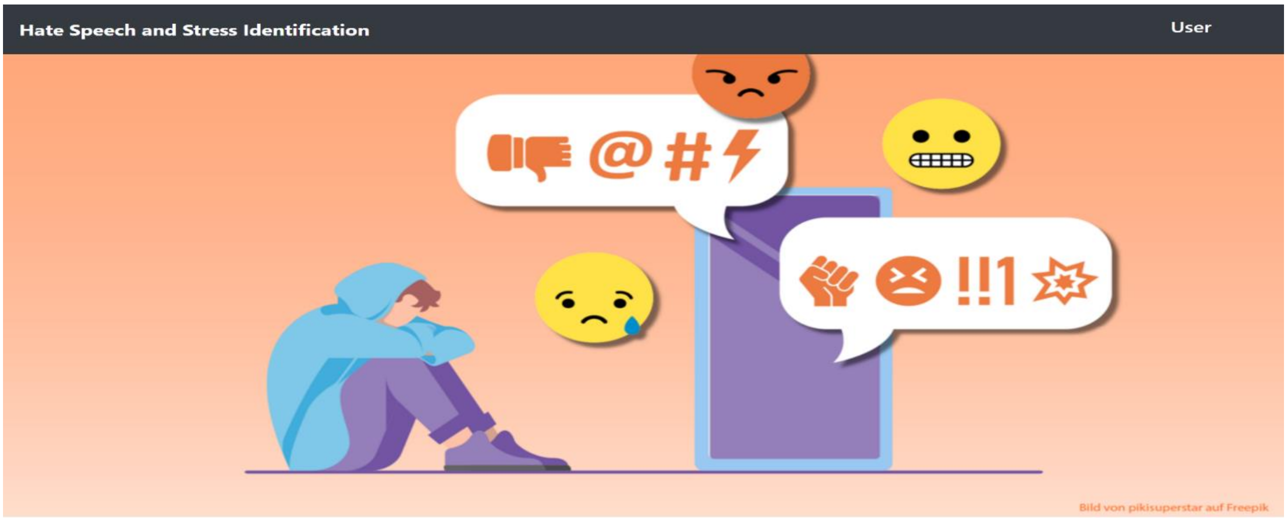
```
39141       normal
Name: label, Length: 39142, dtype: object
(29356, 46121)
vocab => 16101
accuracy:-> 0.929695483343552
              precision    recall  f1-score   support

   depressed       0.88      0.98      0.93      4567
      normal       0.98      0.88      0.93      5219

    accuracy                           0.93      9786
   macro avg       0.93      0.93      0.93      9786
weighted avg       0.94      0.93      0.93      9786
```

```
I cant talk as i am insulted by teammate -> depressed
I am sleeping -> normal
```

Heatmap

This figure shows how hate speech and depressed words are correlated to each other. And above we used about 35000 data sets which are 75% are trained and 25% are used for testing the texts. Using these predicted texts we drown heatmap and same is used for output interface.

The figure 1 shows the interface page of project

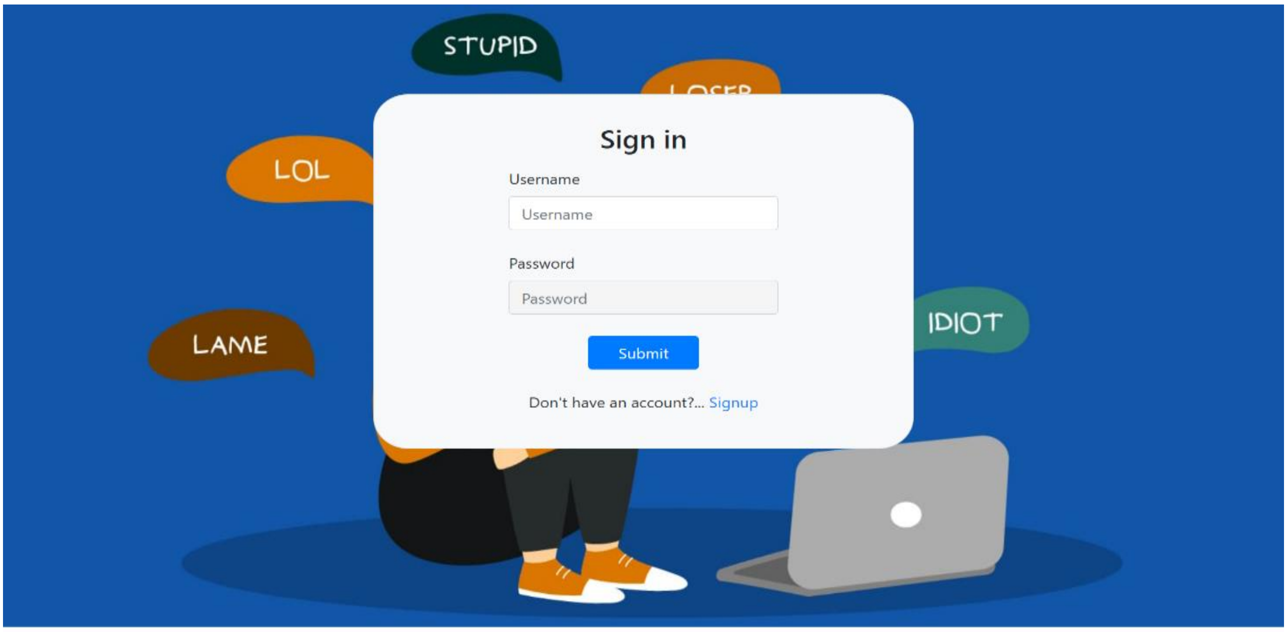


Hate Speech and Stress Identification
User
@#
!!1
Bild von pikisuperstar auf Freepik

Figure 1

In figure 2 we can see the login page of the project which a admin can login through it.

DOI: 10.17148/IJARCCE.2024.134121



Sign in

Username

Username

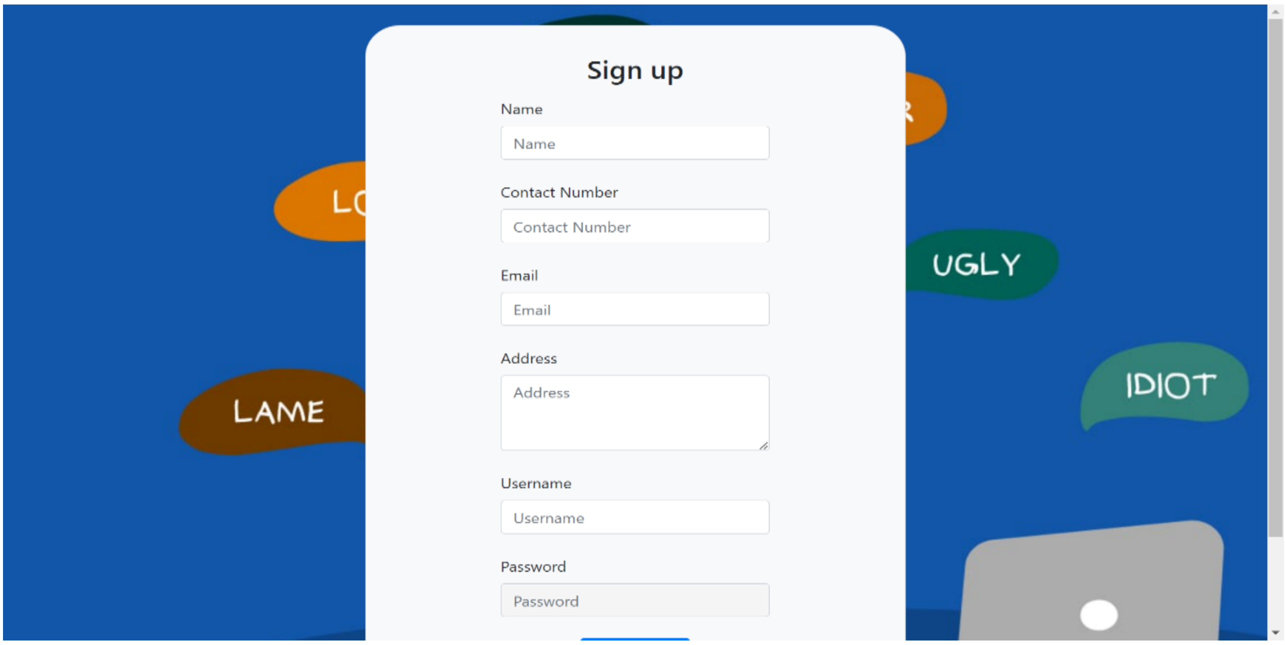
Password

Password

Don't have an account?... Signup

Figure 2

In figure 3 it shows signup page where new user can have their own id and password for his account and he can check the hatespeech words and stress level.



Sign up

Name

Name

Contact Number

Contact Number

Email

Email

Address

Address

Username

Username

Password

Password

Figure 3

In figure 4 we can see the output for the given text which tells that given hatepseech word is commented with stress or not

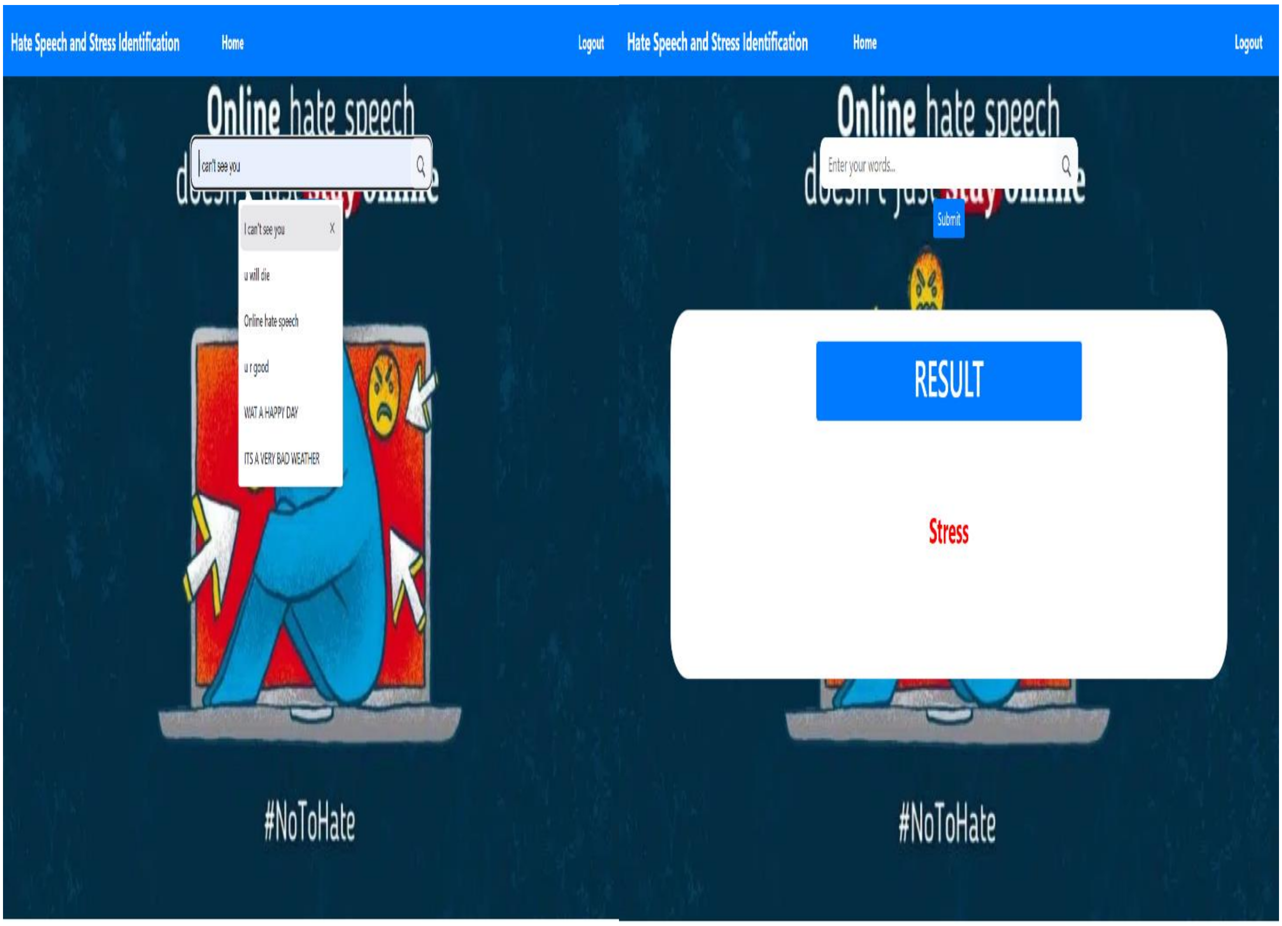


Figure 4

VII. CONCLUSION

The Expected Outcomes Specification is a pivotal aspect of the project, outlining the anticipated results and impacts across multiple facets. In the domain of hate speech detection, the project strives for heightened accuracy by leveraging advanced algorithms and features.

The objective is to surpass conventional benchmarks, enhancing the system's ability to discern and categorize hate speech with precision and recall rates that reflect a substantial improvement. Contextual understanding within hate speech detection is a critical focus, with the anticipation of reducing false positives and negatives through nuanced language comprehension.

The project's cross cultural adaptability outcome aims to navigate diverse cultural contexts effectively, mitigating biases and fostering inclusivity in hate speech detection. Turning to stress detection, the project envisions the effective identification of stress indicators, utilizing linguistic patterns to contribute to user well-being and mental health awareness. The machine learning integration goals center around achieving robust and adaptable models, demonstrating resilience to the dynamic nature of online communication.

VIII. FUTURE WORK

Future work for the project on Social Media-Based Hate Speech and Stress Identification using Machine Learning and Natural Language Processing could involve exploring multimodal analysis incorporating image and video processing, fine-grained classification for nuanced understanding, dynamic adaptation to evolving discourse patterns, multilingual support for global applicability, contextual understanding integrating user demographics and temporal dynamics, real-time monitoring for timely intervention, user-centered design for enhanced usability, ethical considerations addressing privacy and bias, collaborative partnerships with stakeholders, and longitudinal studies to assess long-term impact

DOI: 10.17148/IJARCCE.2024.134121

REFERENCES

[1] Filip Klubicka, Raquel Fernandez “Examining a hate speech corpus for hate speech detection and popularity prediction” arXiv:1805.04661v1 [cs.CL] 12 May 2018.

[2] Raquel Fernandez et al. Published "Hate Speech Corpus Research for Detecting Hate Speech and Predicting Popularity". the Twitter corpus collected by Waseem and Hovy (2016).

[3] Pete Burnap and Matthew L. Williams “Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modelling for Policy and Decision Making .” 1944-2866 # 2015 The Authors. Policy & Internet published by Wiley Periodicals, Inc. on behalf of Policy Studies Organization.

[4] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, “Statistical approaches to concept-level sentiment analysis,” IEEE Intelligent Systems, vol. 28, no. 3, pp. 6–9, 2013 .

[5] Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. ”Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media” .Applied Sciences 10, no. 12: 4180. https://doi.org/10.3390/app10124180.Gaydhani, Aditya, et al. \"Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach.\" arXiv preprint arXiv: 1809.08651 (2018).

[6] Poletto, F., Basile, V., Sanguinetti, M. et al. “Resources and benchmark corpora for hate speech detection: a systematic review”. Lang Resources & Evaluation 55, 477– 523 (2021). https://doi.org/10.1007/s10579-020- 09502-8.

[7] D. G. Haryadi, J. A. Orr, K. Kuck, S. McJames, and D. R. Westen-skow, “Partial co2 rebreathing indirect fick technique for non-invasive measurement of cardiac output.” Journal of clinical monitoring and computing, vol. 16, no. 5-6, pp. 361–74, 2000.

[8] Dmitry D, Oren T, Ari R. “Enhanced sentiment learning using twitter hashtags and smileys”. Coling 2010—23rd International Conference on Computational Linguistics, Proceedings of the Conference. 2. 2010; 241–249.

[9] Krestel R, Fankhauser P. Personalized topic-based tag recommendation. Neurocomputing. 2012;76:61–70. https://doi.org/10.1016/j.neucom.2011.04.034.

[10] Boiy E, Moens MF. A machine learning approach to sentiment analysis in multilingual web texts. Inf Retrieval. 2009;12:526–58. https://doi.org/10.1007/s10791-008-9070-z[N+1]