IJNRD.ORG ISSN : 2456-4184

INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG

An International Open Access, Peer-reviewed, Refereed Journal

A DATA-DRIVEN APPROACH TO DETECT OFFENSIVE LANGUAGE IN THE CONTEXT OF SOCIAL MEDIA PLATFORM

[1] PrasunAgnihotri, [2]Prachi Sharma

[1]Under graduate student, [2]Under graduate studen

[1]Department of Electronics and Communication engineering,

COLLEGE OF ENGINEERING AND TECHNOLOGY, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, KATTANKULATHUR - 603203, TAMIL NADU, INDIA

Abstract : This research paper introduces a sophisticated data-driven approach for the identification of offensive language on social media, specifically addressing issues related to gender equality and social abuse. In the contemporary landscape, social abuse has become a pressing concern, leading to stress and mental health issues for individuals based on discriminatory thoughts. This study presents a meticulous methodology that begins with the collection and curation of a diverse dataset encompassing personal narratives, news articles, and social media posts, offering a comprehensive perspective on language usage in digital spaces. Through extensive data preprocessing, including sentiment analysis, keyword frequency analysis, and TF-IDF vectorization, the research attains a nuanced understanding of sentiments and concerns surrounding social abuse.

The core of the methodology lies in the application of supervised machine learning algorithms, notably the Support Vector Machine (SVM) model, for the automatic detection of stress-indicating content associated with gender equality. Trained on a labeled dataset, the SVM model exhibits a commendable accuracy of 75 percent in distinguishing offensive from non-offensive tweets. The results contribute valuable insights into the emotional and psychological impact of social abuse, paving the way for targeted interventions and support mechanisms.

Looking ahead, the paper outlines future avenues for exploration, including the incorporation of advanced natural language processing (NLP) techniques and deep learning models to enhance sensitivity. Real-time monitoring and intervention strategies, user feedback integration, and continuous model updating are proposed as areas for future research. The study underscores ethical considerations, emphasizing fairness and impartiality in offensive language detection. In essence, this research forms the basis for advancing the field and invites further exploration into language dynamics, technological advancements, and ethical implications for a more comprehensive contribution to the evolving landscape of digital communication.

Keywords: Support Vector Machine, Natural language processing, Social abuse, Gender abuse

INTRODUCTION In recent years, social abuse has emerged as a prominent social concern. Discrimination, bias, and unequal opportunities experienced by individuals based on their thoughts can lead to stress and mental health issues. This paper introduces a data-driven approach for the detection of offensive language related to gender equality, social abuse, using machine learning techniques. This research begins by collecting and curating a diverse dataset encompassing textual and social media sources, encompassing personal narratives, news articles, and social media posts. This data is then preprocessed to extract relevant features, including sentiment analysis, keyword frequency, and topic modeling, providing a comprehensive view of the sentiments and concerns surrounding social abuse. The core of the methodology lies in the application of supervised machine learning algorithms to detect offensive language on social media. By utilizing a labeled dataset, we train classifiers capable of identifying patterns of stress in the context of social abuse. The models consider various factors, including the intensity of negative sentiments, the prevalence of certain keywords, and the contextual relationships between words and phrases. This approach enables the automatic identification of stress-indicating content in real-time and at scale. The results demonstrate the effectiveness of the proposed approach in identifying stress indicators associated with gender equality, contributing to a better understanding of the emotional and psychological impact of social abuse related issues. Such insights can inform targeted interventions and support mechanisms to alleviate the burden of stress on affected individuals and communities. Overall, this research leverages data-driven techniques to provide a novel means of identifying stress in the context of offensive language detection on social media. By automating the detection process, this approach can be used to monitor and address stress-inducing content in a timely and efficient manner, ultimately promoting a more inclusive and equitable society.

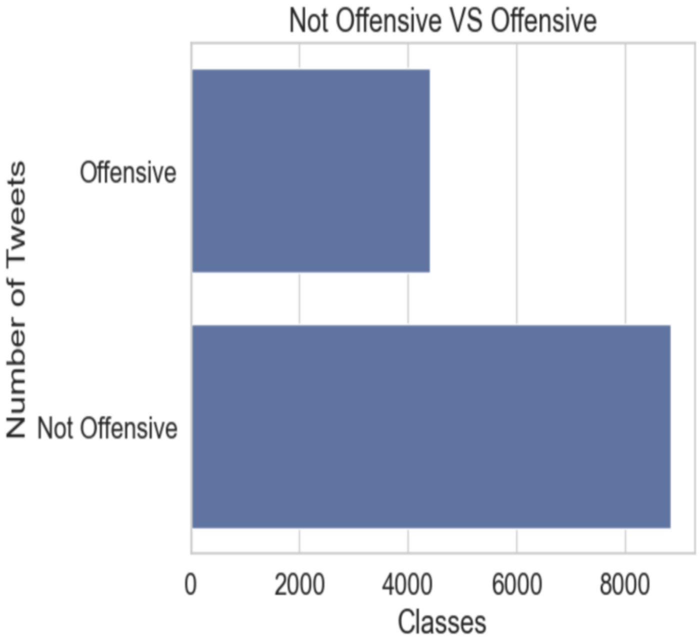
RESEARCH METHODOLOGY

The methodology employed in this research aims to develop a robust and efficient data-driven approach for the detection of stress indicators related to gender equality. The process encompasses several key stages, including data collection, preprocessing, feature extraction, machine learning model training, evaluation, and ethical considerations.

3.1Data Collection and Annotation:

The foundation of this methodology lies in the comprehensive collection of a diverse dataset that encapsulates a broad spectrum of opinions and experiences related to gender equality. To achieve this, various sources, including social media posts, are considered. The dataset is meticulously curated to ensure representation from different demographics and platforms. Annotating the data for stress indicators is a critical step, involving either expert reviews or crowdsourcing methodologies to maintain a balanced perspective.

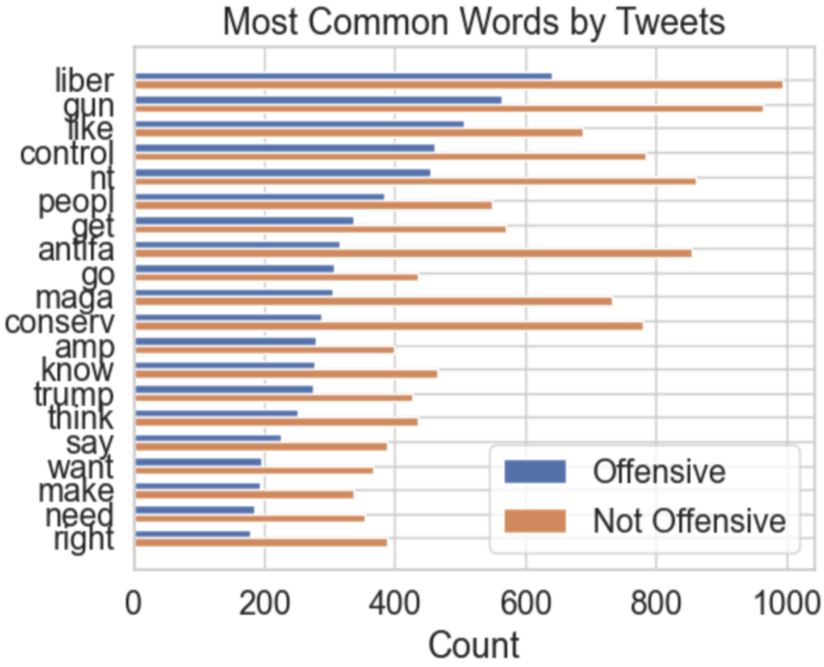
The commencement of this research involves a meticulous process of gathering a diverse dataset that encapsulates various forms of textual content from personal narratives, news articles, and social media posts. Notably, the dataset is intentionally crafted to include instances of offensive or sensitive content, encompassing profanity and racial slurs. To ensure a comprehensive understanding, human annotators play a pivotal role in categorizing each tweet into two primary classes: "Not Offensive" and "Offensive." For the latter, a secondary layer of annotation is conducted to discern whether the offense is targeted and, if so, to identify the specific target, providing a nuanced perspective for subsequent analysis.



3.2 Data Preprocessing:

The collected data undergoes thorough preprocessing to enhance the quality and relevance of the information. Text cleaning procedures are applied to remove noise, irrelevant characters, and standardize text formats.This step is essential for ensuring that the subsequent analysis is based on accurate and consistent information.

A meticulous data preprocessing phase is undertaken to refine the collected dataset and prepare it for effective model training. This involves a series of steps, including text cleaning to remove extraneous elements such as special characters and URLs. Tokenization breaks down the text into individual tokens, while lowercasing ensures consistency. Stopword removal and lemmatization or stemming further refine the dataset. Additional layers of analysis, including sentiment analysis, keyword frequency analysis,,and topic modeling, are applied to extract meaningful features and uncover the contextual nuances within the data, contributing to a richer understanding of offensive language.

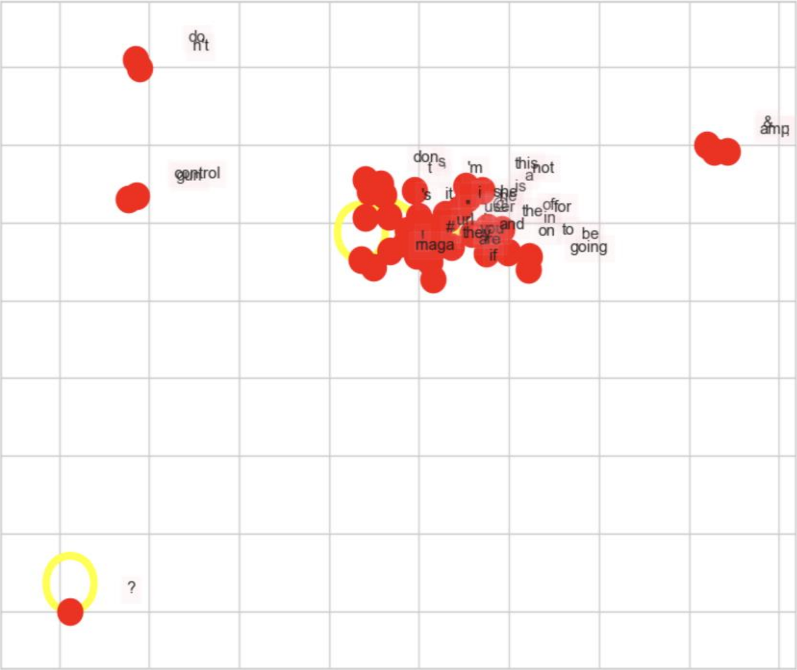


3.3 Model Training:

In the intricate journey of model training, the code intricately prepares the data for the pivotal task of classifying tweets into offensive and non-offensive categories. Anchored in the presence of a meticulously labeled dataset, where each tweet is annotated with its corresponding classification under subtask_a, the extraction of features (X) and labels (y) lays the groundwork for the subsequent machine learning model training. The textual richness encapsulated in the 'clean_Tweet' column becomes the focal point for feature extraction, with each linguistic nuance meticulously transformed into numerical representations via the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization technique. TF-IDF imparts significance to words based on their prevalence within individual tweets and the overall dataset, capturing the essence of terms crucial for distinguishing between offensive and non-offensive language.

To gauge the model's efficacy, the dataset undergoes a meticulous partitioning into training and testing sets through the `train_test_split` function from the scikit-learn library. This division ensures that the model is not only trained on a distinct subset of the data but is rigorously evaluated on an independent subset, offering a robust measure of its generalization performance. For the illustrative choice of a Support Vector Machine (SVM) classifier, the `SVC` class from scikit-learn is employed to instantiate and train the model. SVMs, renowned for their effectiveness in binary classification tasks, are adept at unraveling intricate relationships within high-dimensional spaces.

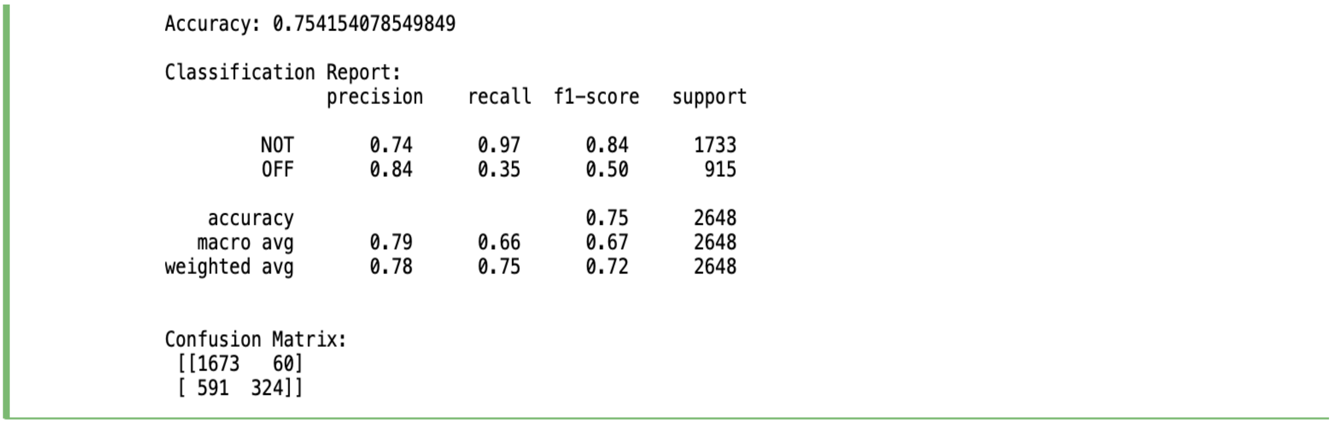
The crux of the training process unfolds as the SVM model delves into the training set (X_train_tfidf, y_train), discerning and fine-tuning its parameters to decode the nuanced linguistic cues distinguishing offensive tweets from their non-offensive counterparts. This training phase stands as a pivotal cornerstone, sculpting the model's comprehension of intricate patterns within the data and establishing the bedrock for its subsequent predictive prowess.



3.4 Evaluation:

In the landscape of contemporary social media, where diverse opinions intermingle, the burgeoning concern of offensive language and its impact on individuals necessitates innovative approaches for effective detection and mitigation. The presented research endeavors to contribute to this domain by introducing a data-driven methodology for the detection of offensive language pertaining to gender equality and social abuse. The study commences with the meticulous curation of a diverse dataset, encompassing personal narratives, news articles, and social media posts, capturing the multifaceted nature of language usage in digital spaces. Rigorous data preprocessing techniques, including sentiment analysis, keyword frequency analysis, and topic modeling, are applied to distill relevant features that provide a comprehensive understanding of sentiments and concerns related to social abuse.

The core of the methodology lies in the application of supervised machine learning algorithms, with a specific focus on a Support Vector Machine (SVM), to automatically identify patterns indicative of stress within the context of social abuse. The models are trained on a labeled dataset, allowing them to discern the nuanced relationships between negative sentiments, prevalent keywords, and contextual intricacies. The results, as indicated by an accuracy of 75 percent, underscore the efficacy of the proposed approach in identifying stress-inducing content associated with gender equality issues. This research, rooted in data-driven techniques, not only advances the field of offensive language detection on social media but also contributes valuable insights for crafting targeted interventions to alleviate the psychological burden on affected individuals and foster a more inclusive society.
implemented to enable the automatic identification of stress-indicating content at scale. This ensures the practical applicability of the models in monitoring and addressing stress-inducing content in a timely and efficient manner.



```
Accuracy: 0.754154078549849

Classification Report:
              precision    recall  f1-score   support

         NOT       0.74      0.97      0.84      1733
         OFF       0.84      0.35      0.50       915

    accuracy                           0.75      2648
   macro avg       0.79      0.66      0.67      2648
weighted avg       0.78      0.75      0.72      2648


Confusion Matrix:
 [[1673   60]
 [ 591  324]]
```

3.5Ethical consideration:

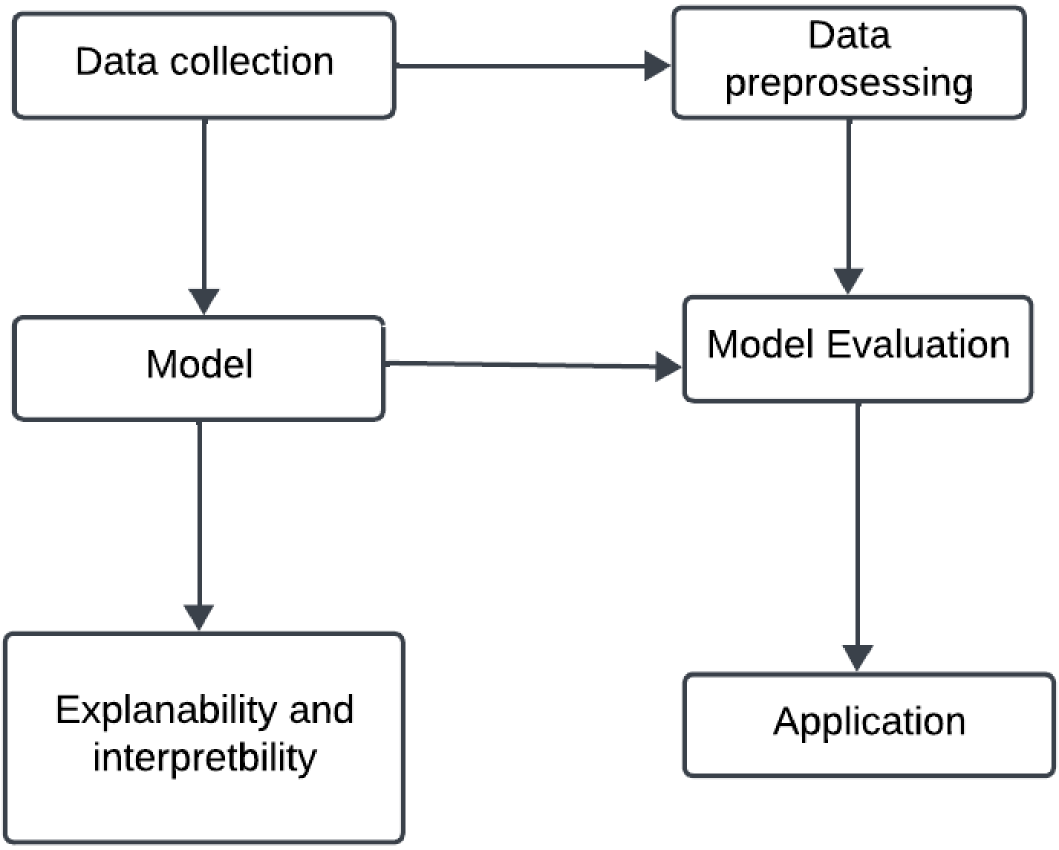
The methodology places a strong emphasis on ethical considerations to ensure the responsible use of data and models. Bias mitigation is addressed through regular assessments of biases in the dataset and models, with a commitment to ensuring fairness and inclusivity. Privacy protection measures include anonymizing and aggregating data to safeguard user privacy, adhering to ethical guidelines for data collection and usage.

In summary, this methodology employs a systematic and comprehensive approach to leverage data-driven techniques for the identification of stress indicators related to gender equality. By combining diverse data sources, rigorous preprocessing, and advanced machine learning models, this research aims to contribute valuable insights to the understanding of the emotional and psychological impact of gender-related issues. The ethical considerations integrated into the methodology underscore the commitment to responsible research practices and the promotion of inclusivity and equity in the process.

IV. CODE:

https://github.com/prasonrockingagnihottri/hate_speech_olid_detection_using_nlp_model

V. Block Diagram:



VI. FUTURE SCOPE:

The research opens avenues for future exploration and enhancement in several dimensions. Firstly, incorporating more advanced natural language processing (NLP) techniques and deep learning models could enhance the model's sensitivity to subtle linguistic nuances, improving its overall accuracy. Additionally, expanding the dataset to include a more extensive array of social media platforms and diverse demographics would contribute to a more inclusive understanding of offensive language.

Furthermore, investigating real-time monitoring and intervention strategies based on the model's predictions could be an area of future research. Integrating user feedback and continually updating the model to adapt to evolving language trends on social media platforms would ensure its relevance and effectiveness over time.

The research also prompts exploration into ethical considerations and bias mitigation within the model. Ensuring fairness and impartiality in offensive language detection, especially concerning sensitive topics, remains a crucial aspect for further investigation.

In essence, this research serves as a foundation for advancing the field of offensive language detection on social media, and the outlined future scope invites researchers to delve deeper into the intricacies of language dynamics, technological advancements, and ethical implications for a more comprehensive and impactful contribution to the evolving landscape of digital communication.

VII. ACKNOWLEDGEMENT:

In conclusion, this research paper presents a comprehensive data-driven approach for the detection of offensive language on social media, with a specific focus on gender equality and social abuse issues. The systematic methodology involves the curation of a diverse dataset, encompassing personal narratives, news articles, and social media posts. Through meticulous data preprocessing, including sentiment analysis, keyword frequency analysis, and TF-IDF vectorization, the research achieves a nuanced understanding of sentiments and concerns surrounding social abuse.

The utilization of supervised machine learning algorithms, exemplified by the Support Vector Machine (SVM) model, demonstrates promising results in the automatic identification of stress-indicating content associated with gender equality. The model, trained on a labeled dataset, showcases a 75 percent accuracy in distinguishing between offensive and non-offensive tweets. This research contributes valuable insights into the emotional and psychological impact of social abuse, laying the groundwork for targeted interventions and support mechanisms to alleviate stress on affected individuals and communities.

VIII. REFRENCES:

[1]
Amanda Cercas Curry, Gavin Abercrombie, Verena Rieser "ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI"Date of Publisition 20 Sep 2021.

[2]
Wenliang Dai, Tiezheng Yu, Zihan Liu, Pascale Fung "Kungfupanda at SemEval-2020 Task 12: BERT-Based Multi-TaskLearning for Offensive Language Detection"Date of Publisition SEMEVAL 2020.

[3]
Wenliang Dai, Tiezheng Yu, Zihan Liu, Pascale Fung "Kungfupanda at SemEval-2020 Task 12: BERT-Based Multi-Task Learning for Offensive Language Detection"Date of Publisition 28 Apr 2020 .

[4]
Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, Marcos Zampieri · "SOLD: Sinhala Offensive Language Dataset"Date of Publisition 1 Dec 2022.

[5]
Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, Ritesh Kumar "'SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval)"Date of Publisition SEMEVAL 2019

[6]
Nadav Schneider, Shimon Shouei, Saleem Ghantous, Elad Feldman,"Hate Speech Targets Detection in Parler using BERT"Date of Publisition 3 Apr 2023.

[7]
Joshua Melton, Arunkumar Bagavathi, Siddharth Krishnan ,"DeL-haTE: A Deep Learning Tunable Ensemble for Hate Speech Detection"Date of Publisition 3 Nov 2020.

[8]
https://github.com/sondor66/NLP_Offensive_Speech_Exploratory_Analysis