ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/349-358
Sphoorti Kadaveru S N *et. al.*, / International Journal of Engineering & Science Research

# OFFENSIVE LANGUAGE DETECTION USINGTEXT CLASSIFICATION

**AVS Radhika[1], Sphoorti Kadaveru,[2] Srujana Kondoju**
**[1]Associate Professor, Department Of Cse, Bhoj Reddy Engineering College For Women, India.**
**[2,3]B. Tech Students, Department Of Cse, Bhoj Reddy Engineering College For Women, India.**

**ABSTRACT**
*There is a concerning rise of various social platforms. Such language might bully or hurt the feelings of an individual or a community. Recently, the research community has investigated and developed different supervised approaches and training datasets to detect or prevent offensive monologues or dialogues automatically. In this study, we propose a model for text classification consisting of modular cleaning phase and tokenizer, three embedding methods, and eight classifiers. Our experiments show a promising result for detection of offensive language on our dataset obtained from Twitter.*
*Considering hyperparameter optimization, three methods of AdaBoost, SVM and MLP had highest average of F1-score on popular embedding method of TF-IDF. Index Terms— offensive language detection, social media, machine learning, text mining. This paper reviews text classification methods for offensive language detection in online platforms. It covers algorithms like Naive Bayes, SVMs, and neural networks, along with feature engineering techniques and evaluation metrics. Insights into current research and future directions are provided.*

## 1-INTRODUCTION

Offensive language detection is a critical component in various digital platforms, serving to maintain a respectful and inclusive online *offensive language on the content generated by the* environment. In today's interconnected world,where communication predominantly occurs through digital mediums, the impact of offensive language can be widespread and harmful.

Firstly, offensive language detection promotes the well-being and mental health of users. Experiencing or witnessing offensive language online can lead to feelings of discomfort, anxiety, and even trauma, particularly for marginalized groups. By swiftly identifying and removing such language, platforms create safer spaces for users to engage in dialogue without fear of harassment or discrimination.

Moreover, offensive language detection reinforces community standards and values. Digital platforms often have guidelines regarding appropriate conduct, and detecting and addressing offensive language helps uphold these standards. It sends a clear message that hate speech, derogatory remarks, and other forms of offensive language are not tolerated, fostering a culture of respect and civility.

Additionally, offensive language detection plays a crucial role in preventing the spread of misinformation and harmful ideologies. Hate speech and discriminatory language can perpetuate stereotypes, fuel division, and incite violence. By identifying and filtering out such content, platforms mitigate the risk of it gaining traction and causing real-world harm.

Furthermore, offensive language detection is essential for brand reputation and user retention. Platforms that fail to address offensive content risk alienating users and damaging their reputation. By demonstrating a commitment to fostering a positive paramount importance in the digital age. It safeguards the well-being of users, reinforces community standards, prevents the spreadof harmful ideologies, and protects brand reputation. By investing in robust detection mechanisms and proactive moderation, digital platforms can create safer, more welcoming spaces for all users to connect, communicate, and collaborate.

## 2-LITERATURE REVIEW

***S. Khorshidi, G. Mohler, and J. G. Carter,*** "Assessing gan-based approaches for generative modelling ofcrime text reports,"

Analysis and modeling of crime text report data has important applications, including refinement of crime classifications, clustering of documents, and feature extraction for spatio-temporal forecasts. Having better neural network representations of crime text datamay facilitate all of these tasks. This paper evaluates the ability of generative adversarial network models to represent crime text data and generate realistic crime reports. We compare four state of the art GAN algorithms in terms of quantitative metrics such as coherence, embedding similarity, negative log-likelihood, and qualitatively based on inspection of generated text. We discuss current challenges with crime text representationand directions for future research.

***O. Jafari, P. Nagarkar, B. Thatte, and C. Ingram,*** "Satellitener An effective named entity recognition modelfor the satellite domain,"

Nowadays, large amounts of data is generated daily. Textual data is generated by news articles, social media such as Twitter, Wikipedia, etc. Managing and inclusive online community, platforms can attract and retain a diverse user base, driving engagement and growth.

In conclusion, offensive language detection is of these large data and extracting useful information from them is an important task that can be achieved using Natural Language Processing (NLP). NLP is an artificial intelligence domain dedicated to processing and analysing human languages. NLP includes many subdomains such as Named Entity Recognition (NER), Entity Linking, Sentiment Analysis, Text Summarization, Topic Modelling, and Speech Processing.

***S. Zhang, O. Jafari, and P. Nagarkar,*** "A survey on machine learning techniques for auto labelling of video, audio, and text data,"

Machine learning has been utilized to perform tasks in many different domains such as classification, object detection, image segmentation and natural language analysis. Data labelling has always been one of the most important tasks in machine learning. However, labelling large amounts of data increases the monetary cost in machine learning. As a result, researchers started to focus on reducing data annotation and labelling costs. Transfer learning was designed and widely used as an efficient approach that can reasonably reduce the negative impact of limited data, which in turn, reduces the data preparation cost. Even transferring previous knowledge from a source domain reduces the amount of data needed in a target domain. However, large amounts of annotated data are still demanded to build robust models and improve the prediction accuracy of the model. Therefore, researchers started to pay more attention on auto annotation and labelling. In this survey paper, we provide a review of previous techniques that focuses on optimizeddata annotation and labelling for video, audio, and text data.

ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/349-358
Sphoorti Kadaveru S N *et. al.*, / International Journal of Engineering & Science Research

***NektariaPotha and ManolisMaragoudakis,*** "Cyberbullying detectionusing time series modeling"

Cyber bullying is a new phenomenon resulting from the advance of new communication technologies including the Internet, cell phones and Personal Digital Assistants. It is a challenging bullying problem occurring in a new territory. Online bullying can be particularly damaging and upsetting because it's usually anonymous or hard to trace. In this paper, the proposed method is utilizing a dataset of real world conversations (i.e. Pairs of questions and answers between cyber predator and the victim), in which each predator question is manually annotated in terms of severity using a numeric label. We approach the issue as a sequential data modelling approach, in which the predator's questions are formulated using a Singular Value Decomposition representation. The motivation of this procedure isto study the accuracy of predicting the level of cyber bullying attack using classification methods and also to examine potential patterns between the lingusticstyle of each predator. More specifically, unlike previous approaches that consider a fixed window of a cyber-predator's questions within a dialogue, we exploit the whole question set and model it as a signal, whose magnitude depends on the degreeof bullying content. Using feature weighting and dimensionality reduction techniques, each signal is straightforwardly parsed by a neural network that forecasts the level of insult within a question given a window between two and three previous questions. Throughout the time series modeling experiments, an interesting discovery was made. By applying SVD on the time series data and taking into account the second dimension (since the first is usually modeling trivial dependencies between instances and attributes) we observed that its plot was very similar to the plot of the class attribute. By applying a Dynamic Time Warping algorithm, the similarity of the aforementioned signals was proved to exist, providing an immediate indicator for the severity of cyber bullying within a give

.

**3-METHODOLOGY**


***Figure 1.1 Software Development Methodology***

**Agile Model**

The Agile model is an iterative approach to software development that emphasizes flexibility, collaboration, and customer feedback. Unlike the Waterfall model, which follows a linear sequence of phases, Agile breaks development into small, manageable increments called iterations, usually lasting one to four weeks. Each iteration involves planning, development, testing, and review, enabling teams to deliver functional components regularly and adapt quickly to changes. Collaboration is key, with cross-functional teams working closely together and holding daily stand-up meetings, sprint planning sessions, and retrospectives. Customer feedback is integral, with regular demonstrations at the end of each iteration allowing stakeholders to provide input that can be incorporated into the next cycle. This approach ensures the product continuously evolves to meet user needs, leading to higher-quality software.

•

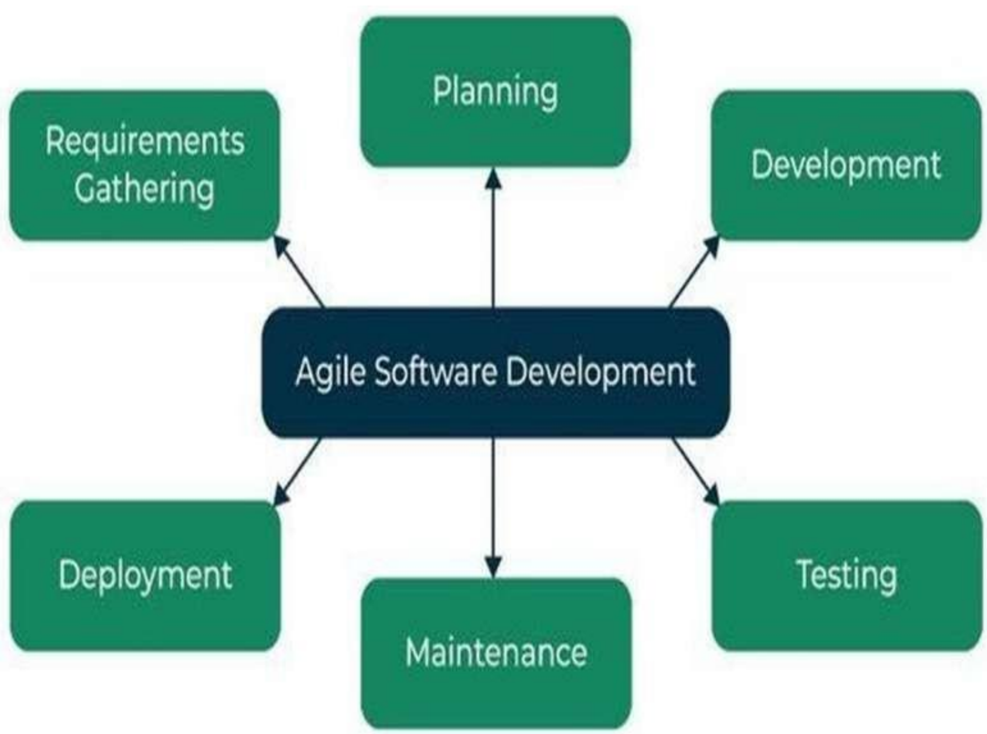



***Figure Agile Development Model***

ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/349-358
Sphoorti Kadaveru S N *et. al.*, / International Journal of Engineering & Science Research

**Waterfall Model**

The Waterfall Model is a traditional linear sequential approach to software development. It divides the software development process into distinct phases, with each phase dependent on thedeliverables of the previous phase. Here's an overview of its key features, advantages, and disadvantages:

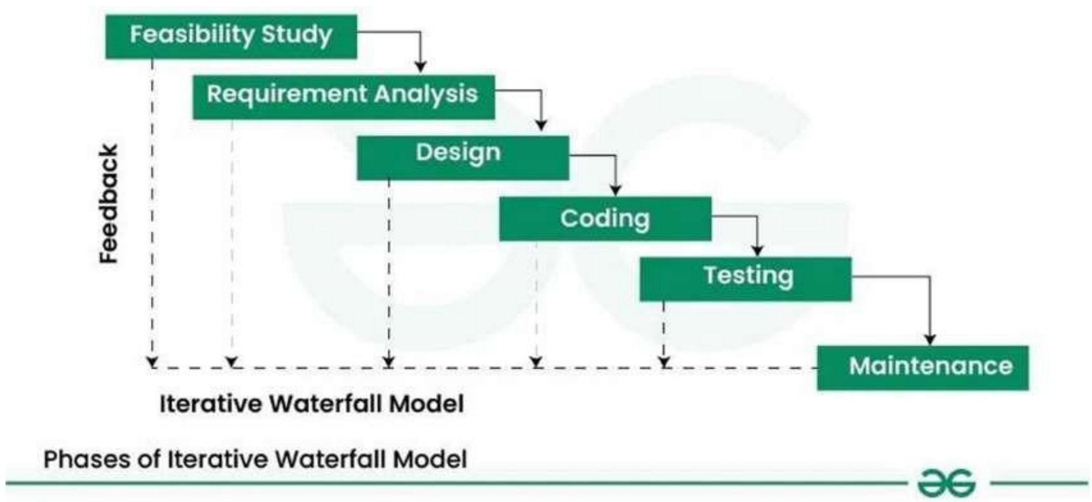
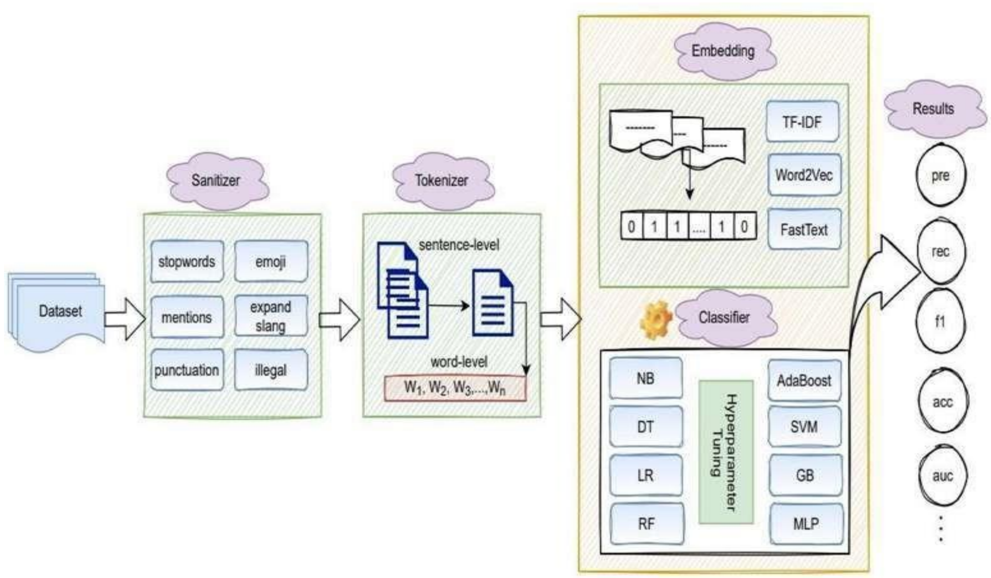



*Fig 1.3 Waterfall Development Model*

**SYSTEM ARCHITECTURE**

The system architecture of this projects shows the flow of the control through thesystem. It also shows the hardware and the software required for the execution of the program. The architecture Diagram is as follows

*Fig 2.1 System architecture*



ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/349-358
**Sphoorti Kadaveru S N** ***et. al.,*** **/ International Journal of Engineering & Science Research**

**4-Implementation Module**

Description

- User
- Admin
- Data Processing
- Machine Learning

**Testing Strategies and Methodologies**

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Testing is an important aspect of any software development project. It ensures that the software is functioning as expected and meets the requirements of the users. There aredifferent testing strategies and methodologies that can be used to test software.

**Testing Strategy**

For our project, the testing strategy will encompass a comprehensive approach to ensurethe functionality, performance, and security of the software. The testing phases will include

- **Unit Testing** This phase will focus on testing the smallest units of the system. We will utilize testing frameworks like Jest for unit tests on individual components ofthe AI algorithms.
- **Integration Testing** Integration testing will examine the interactions between different modules of the system. In our project, integration tests will assess the seamless integration between AI models, databases, and external APIs.
- **System Testing** System testing will evaluate the overall functionality of the system, including user interfaces and backend processes. Manual testing will validate the user experience, while automated testing tools like Cypress will ensure the robustness of the system.
- **Acceptance Testing** This phase will verify that the system meets stakeholder requirements. Acceptance tests will be designed to cover various scenarios, including positive and negative test cases, ensuring alignment with stakeholder expectations. Additionally, user feedback will play a crucial role in refining the system during acceptance testing.

***fig 3.1 Showing levels of testing strategies…***

## 5-RESULTS

**Main Window/ Home Page**

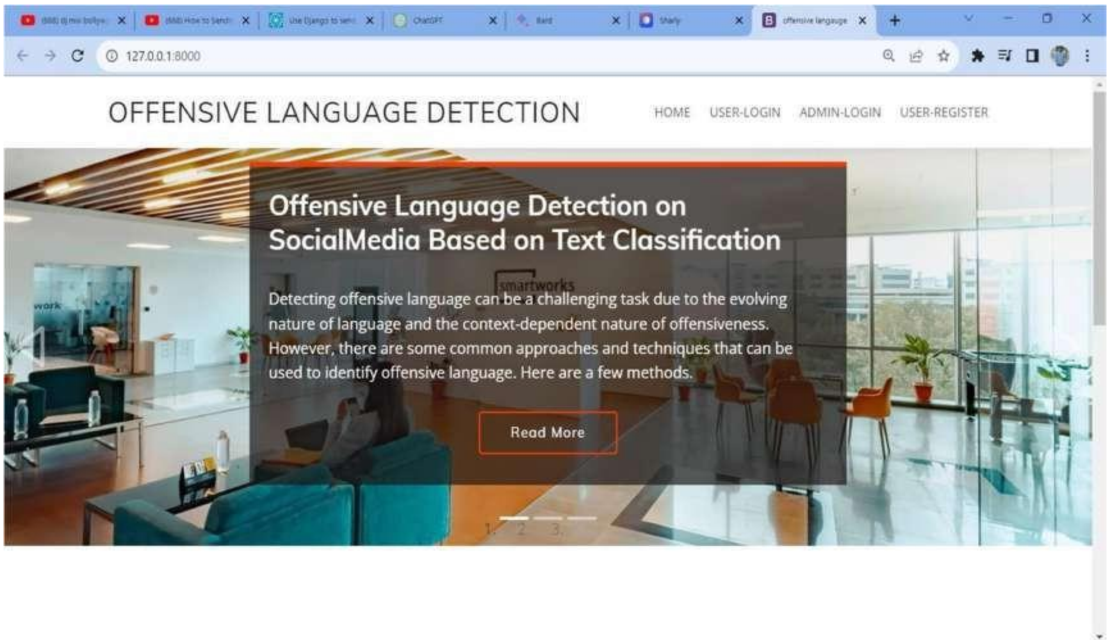



*fig 1 Showing Home Page…*


*fig 2 Profile Creation…*


*fig 3 Admin login*

ISSN 2277-2685
IJESR/June. 2025/ Vol-15/Issue-3s/349-358
Sphoorti Kadaveru S N *et. al.*, / International Journal of Engineering & Science Research


*fig 4 User login…*

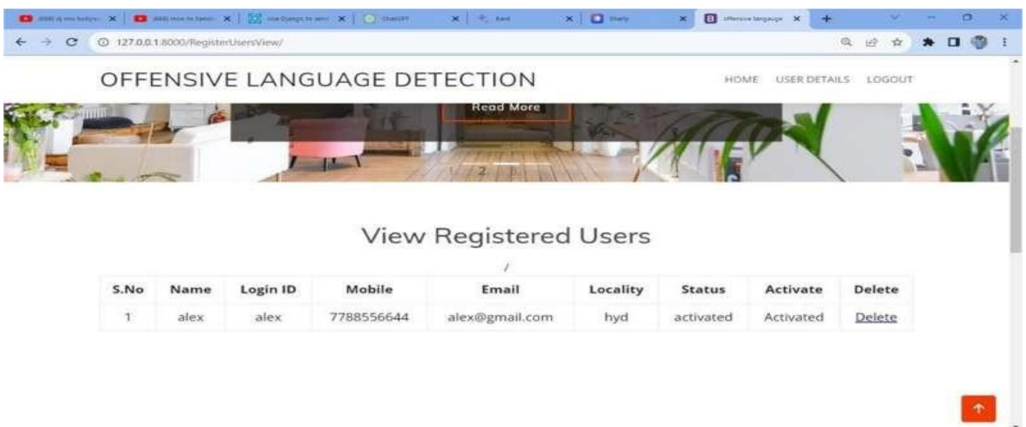



***Fig 5 Registered Users Details…***

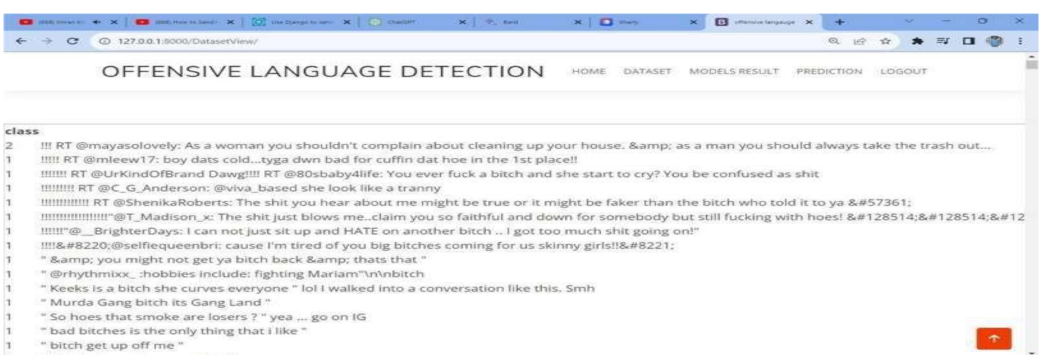
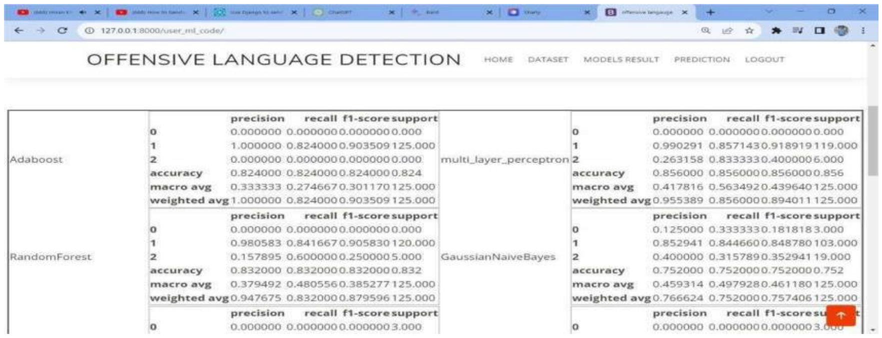



***Fig 6 Dataset Used for training…***



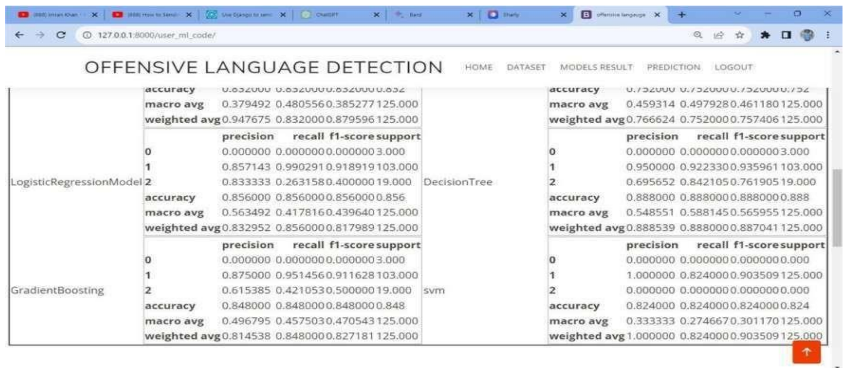




***fig 7 Showing Results of Algorithms…***



***Fig 4.8 Showing Result/ Prediction…***

## 6-CONCLUSION

In this work, we propose a modular text classification pipeline on social media datasets focusing on Twitter. Our proposed approach is to leverage a modular development that allows easy use for combining different text classification components. This paper's maincontribution is that it presents a new modular text classification pipeline to facilitate benchmarking by conducting a detailed analytical study of the best-performing approaches, features, and embeddings reported by the state-of- the-art.

In conclusion, offensive language detection software stands as an indispensable tool in our increasingly digital world. Its significance lies not only in safeguarding against the proliferation of harmful language but also in fostering inclusive and respectful online environments. By automating the identification and moderation of offensive content, such software not only protects individuals from harassment and discrimination but also upholds the values of diversity and equality. As we continue to navigate the complexities of online communication, the importance of robust offensive language detection software cannot beoverstated, serving as a vital guardian of digital civility and human dignity.

## REFERENCES

[1] N. Sabetpour, A. Kulkarni, and Q. Li, "OptSLA an optimization-based approach for sequential label aggregation," in Findings of the Association for Computational Linguistics EMNLP 2020. Online Association for Computational Nov. 2020, pp. 1335–1340.

[2] N. Sabetpour, A. Kulkarni, S. Xie, and Q. Li, "Truth discovery in sequence labels from crowds," 2021

[3] S. Khorshidi, G. Mohler, and J. G.Carter, "Assessing gan-based approaches for generative modeling of crime text reports," in 2020 IEEE International Conference on Intelligence and Security Informatics (ISI). IEEE, 2020

[4] O. Jafari, P. Nagarkar, B. Thatte, and C. Ingram, "Satellitener An effective named entity recognition model for the satellite domain," in Proceedings of the KMIS 2020, vol. 3, 2020, pp.
100–107

[5] S. Zhang, O. Jafari, and P. Nagarkar, "A survey on machine learning techniques for auto labeling of video, audio, and text data," arXiv preprint arXiv 2109.03784, 2021.

[6] M. Saadati, J. Nelson, A. Curtin, L. Wang, and H. Ayaz, "Application of recurrent convolutional neural networks for mental
workload assessment using functionalnearinfrared spectroscopy," in Advances in Neuroergonomics and Cognitive Engineering, H. Ayaz, U. Asgher, and L. Paletta, Eds. Cham Springer International Publishing, 2021, pp. 106–113.

[7] N. Kalantari, D. Liao, and V. G. Motti, "Characterizing the online discourse in twitter Users' reaction to misinformation around covid-19 in twitter," in 2021 IEEE International Conference
on Big Data (Big Data), 2021, pp. 4371–4380.

[8] A. Esmaeilzadeh, M. Heidari, R. Abdolazimi, P. Hajibabaee, and M. Malekzadeh, "Efficient large scale nlp feature engineering with
apache spark," in 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2022.

[9] R. Abdolazimi, M. Heidari, A. Esmaeilzadeh, and H. Naderi, "Mapreduce preprocess of big graphs for rapid connected components detection," in 2022 IEEE 12th Annual Computing and Communication Workshop and