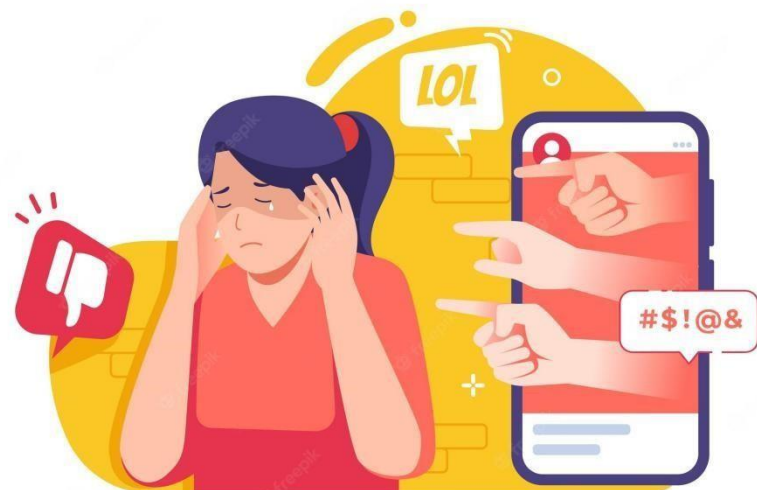




Cyberbullying prevention by analyzing Social Media posts and comments using CNN AND LSTM

1. ABSTRACT

To protect users from being exposed to offensive language on online forums or social media sites, companies have started flagging comments and blocking users who are found guilty of using unpleasant language. Several machine learning models have been developed and deployed to filter out unruly language and protect internet users from becoming victims of online harassment and cyberbullying. The researchers also found that experiencing cyberbullying increased the risk of thoughts of suicide. Victims share Negative feelings that express hardship, thoughts of death, and self-harm are widespread on social media. Therefore, using social media to detect and identify suicidal ideation will help provide



proper intervention that will eventually dissuade others from self-harming and committing suicide and prevent the spread of suicidal ideations on social media. This project describes a way to classify toxic comments on Instagram posts by flagging users who post toxic comments and then tracking further activities to check whether they are making any suicidal posts.

2. Keywords

Suicidal ideation detection, mental health, social content, feature engineering, deep learning, Toxic comments classification, Long Short Term Memory model (LSTM), Convolutional Neural Network (ConvNet/CNN)

3. Introduction

The use of social media platforms and microblogging websites for interpersonal and group communication has grown rapidly. People use comments and feedback on these platforms to communicate their views, ideas, and opinions as well as to express their emotions [1]. In a recent survey done by McAfee Corp (July 2022) [9], Around 85 per cent children in India have reported being cyberbullied and it is the highest in the world. From 2.4 billion in 2014 to 3.4 billion, 4 billion, and 4.4 billion in 2016, 2017, and June 2019, respectively, there been a gradual increase in internet users every year [2]. The number of internet users will reach 4,648 billion in May 2020 [3]. These people have a common space on social media platforms where they may express their viewpoints and engage in discussion. However, issues develop when social media platforms are used as the venue for arguments and disputes that involve the use of toxic (also known as unpleasant, disrespectful, and offensive) comments. Online comments are fraught with dangers like toxicity, bogus news, cyberbullying, and online harassment [4]. Unfortunately, users now experience a variety of psychological issues, including depression, frustration, and even suicidal thoughts as a result of these toxic comments, which harm the reputation of social media platforms. [1].

To avoid the aforementioned problems and keep online conversations stable, toxic comment classification is crucial [5]. Online harassment, personal attacks, and bullying are all examples of toxic remarks. Numerous examples of police arrests due to abusive or defamatory comments on personal sites have occurred over the past few years, according to the 314 International Journal for Modern Trends in Science and Technology [6], [7]. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions. These toxic comments have become a serious issue that affects the reputation of social platforms and causes different psychological problems for users, such as depression,frustration, and even suicidal thoughts. The researchers also found that experiencing cyberbullying increased the risk of thoughts of suicide. Some online messages are really depressing and lead to unpleasant phenomena like cyberstalking and cyberbullying. Since such poor information frequently engages in some type of social cruelty, the consequences can be serious and dangerous,resulting in gossip or even mental harm. Cyberbullying and suicide are related, according to research [10]. When exposed to too many bad messages or experiences, victims may experience depression and desperation. Worse yet, somemay even attempt suicide. Worldwide, suicidal thought affects people of all ages as a result of shock, fury, guilt, andother depressive or anxious feelings.

This project aims to build a predictive web app to detect toxicity in users instagram posts and detect suicidal intent in further instagram posts of victim. To further highlight the applicability of our webapp, we suggest motivational videos for victim.

4. Literature Review

	PAPER 1	PAPER 2	PAPER 3	PAPER 4	PAPER 5
Datasets/corpora	<u>Kaggle</u> <u>Wikipedia's talk</u> <u>page edits dataset</u>	wikipedia_toxicity_ subtypes		Text Data: a) Reddit: b) Twitter: c) Reach Out: 2) EHR: 3) Mental Disorders	
used data			<ul style="list-style-type: none"> sample of 500 titles of posts published in Reddit's Suicide Watch forum Twitter dataset for suicide risk assessment 		www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
sources used	<u>Kaggle</u>	Wikipedia	<ul style="list-style-type: none"> Reddit Twitter 	<ul style="list-style-type: none"> Reddit Twitter 	Kaggle
scope	To identify the toxic comments on social media platforms	To create an online interface where we would be able to identify the toxicity level in the given phrase or sentence and classify them into their order of toxicity	<p>This paper aimed to describe an approach for the suicide risk assessment of Spanish-speaking users on social media. We aimed to explore</p> <p>behavioral, relational, and multimodal data extracted from multiple social platforms and develop machine learning models</p>	<p>This article is the first survey that comprehensively introduces and discusses current suicidal ideation detection (SID) methods including clinical methods based on the interaction between social workers or experts and the targeted individuals and machine learning techniques</p> <p>with feature</p>	Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification.

Applications of Algorithms used	<ul style="list-style-type: none"> Sentiment analysis for employees classification and prediction of diabetes mellitus using soft voting classifier Predicting death risk analysis in fully vaccinated people 	<ul style="list-style-type: none"> Trending Topic Classification for Single-Label Using Multinomial Naive Bayes (MNB) and Multi-Label Using K-Nearest Neighbors (KNN) Personality classification based on Twitter text using Naive Bayes, KNN and SVM 	<ul style="list-style-type: none"> Decoding Facial Recognition Analyzing Documents Collecting Historic and Environmental Elements Understanding Climate 	<ul style="list-style-type: none"> Prediction problems. Language Modelling and Generating Text. Machine Translation. Speech Recognition. Generating Image Descriptions. Video Tagging. 	<ul style="list-style-type: none"> Robot control. Time series prediction. Speech recognition. Rhythm learning. Music composition. Grammar learning
	important existing algorithms / approaches used	SVM, RF, GBM, and LR, the proposed RVVC and RNN deep learning models	Binary relevance method with Multinomial Naïve Bayes and Support vector classifier	BoW model trained with 1 to 5 grams deep learning model defined by a CNN architecture	logistic regression BR Method with Multinomial Naive Bayes classifiers BR Method with SVM classifier
supervised or unsupervised	supervised	supervised	supervised	supervised	supervised

similarity measures used	RVVC outperforms all other individual models when TF-IDF features are used with SMOTE balanced dataset and achieves an accuracy of 0.97	The outcomes for the algorithms were as follows, if we compare both hamming losses, we could conclude that Naïve Bayes has a hamming loss of 3.6 and an accuracy of 87.6 whereas the hamming loss for SVM is 4.36 and the accuracy is 88.16. This gives us	The types of attributes analyzed are significant for detecting users at risk, and their combination outperforms the results provided by generic, exclusively text-based baseline models. After evaluating the	1: Convolutional Neural Networks (CNN) 2: Recurrent Neural Network (RNN)	Our LSTM and CNN models performed the best for word-level binary and multi-label classification. Our CNN models fared the best for character-level binary classification, but overall our word-level models beat our character-level
--------------------------	---	--	---	---	--

		a brief insight to understand the optimal algorithm that can be utilized for ordering toxic comments	contribution of image-based predictive models, we believe that our results can be improved by enhancing the models based on textual and relational features. These methods can be extended and applied to different use cases related to other mental disorders.		models. As indicated by Nguyen and Nguyen, we also tried stacking a character-level CNN with a bidirectional LSTM, but we were unable to replicate the excellent accuracy metrics they achieved in their work.
the evaluation metrics specified in the evaluation methodologies section	Here they evaluate the performance of machine learning models in terms of accuracy, precision, recall, and F1 score	From the results we can conclude that taking hamming loss as a measure of identifying the optimal algorithm to classify toxic comments we can say that Binary Relevance method with Multinomial Naive Bayes is an efficient algorithm that serves our purpose and has a hamming loss of 3.6 as compared to the hamming loss of SVM with a score of 4.36	<ul style="list-style-type: none"> ● Precision ● F1 score ● Accuracy ● Area Under Curve (AUC) 	SVM artificial neural networks (ANN) conditional random field (CRF)	<ul style="list-style-type: none"> ● Accuracy ● Precision ● Recall ● F1-score

TABLE-1

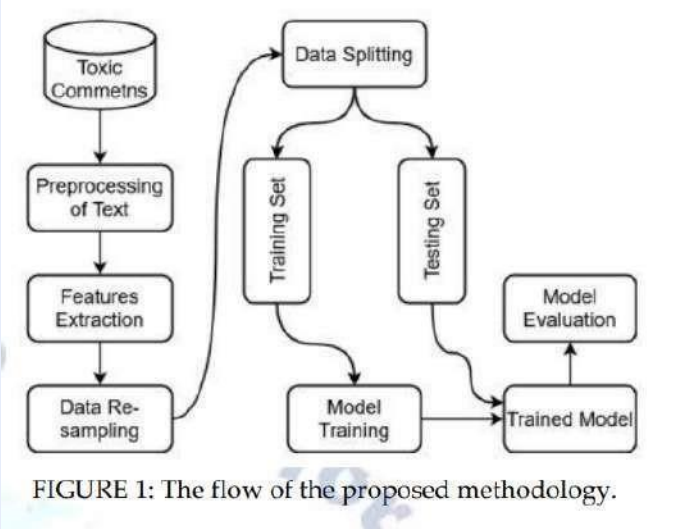
<i>Team No.</i>	Curiosity Rover
-----------------	-----------------

<p><i>Paper Title</i></p>	<ul style="list-style-type: none"> ● A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning ● An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding ● Imbalanced Text Features for Toxic Comments Classification ● Identification and Classification of Toxic Comment Using Machine Learning Methods ● Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis ● Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications ● Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods ● Machine learning methods for toxic comment classification: a systematic review ● Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification ● A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning ● A Comparative Analysis on Suicidal Ideation Detection Using NLP, Machine, and Deep Learning
<p><i>Citation (IEEE style)</i></p>	<ul style="list-style-type: none"> ● Basha, S & Reddy, T & Ijmtst, Editor. (2022). Imbalanced Text Features for Toxic Comments Classification. International Journal for Modern Trends in Science and Technology. 8. 313-317. 10.46501/IJMTST0801054. ● P.Vidyullatha¹ , Satya Narayan Padhy¹ ,Javvaji Geetha Priya² , Kakarlapudi Srija³ ,Sri Satyanjani Koppiseti⁴ ¹Associate Professor, Dept. of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, A.P., India ^{1,2,3,4}Vth year B.Tech Student, Dept. of Computer Science

	<p>and Engineering, Koneru Lakshmaiah Education Foundation, A.P., India</p> <ul style="list-style-type: none"> ● Ramírez-Cifuentes, Diana, et al. "Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis." <i>Journal of medical internet research</i> 22.7 (2020): e17758. ● Ji, Shaoxiong, et al. "Suicidal ideation detection: A review of machine learning methods and applications." <i>IEEE Transactions on Computational Social Systems</i> 8.1 (2020): 214-226. ● Abdulsalam, Asma, and Areej Alhothali. "Suicidal Ideation Detection on Social Media: A Review of Machine Learning Methods." <i>arXiv preprint arXiv:2201.10515</i> (2022). ● Andročec, Darko. "Machine learning methods for toxic comment classification: a systematic review." <i>Acta Universitatis Sapientiae, Informatica</i> 12.2 (2020): 205-216. ● Maslej-Krešňáková, Viera, et al. "Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification." <i>Applied Sciences</i> 10.23 (2020): 8631. ● Haque, Rezaul, et al. "A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning." <i>Technologies</i> 10.3 (2022): 57. ● Alsharef, Ahmad, et al. "An automated toxicity classification on social media using LSTM and word embedding." <i>Computational Intelligence and Neuroscience</i> 2022 (2022).
--	--

TABLE-2 (ARTICLE 1)	
Problem answered in this paper	To identify the toxic comments on social media platforms

Detailed description about the problem	Social media platforms and microblogging websites have gained accelerated popularity during the past few years. These platforms are used for expressing views and opinions about products, personalities, and events. Often during discussions and debates, fights take place on social media platforms which involves using rude, disrespectful, and hateful comments called toxic comments.
	The identification of toxic comments has been regarded as an essential element for social media platforms.
Why that problem is chosen in this paper? Scope of the problem and solution (Refer Introduction)	platforms and cause different psychological problems for users, such as depression, frustration, and even suicidal thoughts . Toxic comment classification is very important to overcome the above-mentioned issues and maintain stability in online debate
History of the problem. (Refer Introduction)	Toxic comments can be considered as a personal attack, online harassment, and bullying behaviors. Over the past few years, several cases of police arrests happened where police arrested many individuals due to the abusive or negative content on personal pages. So a framework that can detect toxic comments and prevent publishing is of significant importance. As a result, several approaches have been introduced for the automatic detection of toxic comments using machine learning algorithms
List of the related/similar problems (Refer Related work) – Describe each with proposed solutions	
What is the proposed solution in this paper for the problem chosen? (Refer Proposed work) (5-8 lines)	The proposed approach is called regression vector voting classifier (RVVC) and combines these models using soft voting criteria . The soft voting criteria ensure that the class with a high predicted probability by two classifiers will be considered as the final prediction.

<p>Architecture of the proposed solution. (Refer proposed work) Diagram</p>	 <p>FIGURE 1: The flow of the proposed methodology.</p>
<p>Name of the approach as stated by the authors</p>	<p>Regression vector voting classifier</p>
<p>(if not, you try to give a name based on the concepts used)</p>	
<p>List of existing algorithms used by the authors to complete the proposed work. (1-2 lines for each algorithm)</p>	<p>SVM, RF, GBM, and LR, the proposed RVVC and RNN deep learning models</p>
<p>List of datasets used. (Refer experimental evaluation/result discussion)</p>	<p>wikipedia_toxicity_subtypes</p>
<p>References/links to each of the dataset used in this paper (in IEEE style)</p>	<p>.https://www.tensorflow.org/datasets/catalog/wikipedia_toxicity_subtypes</p>

<p>List of method(s)/metrics used to evaluate the proposed approach. (Refer experimental evaluation/result discussion)</p>	<p>The proposed approach is called regression vector voting classifier (RVVC) and combines these models using soft voting The soft voting criteria ensure that the class with a high predicted probability by two classifiers will be considered as the final prediction</p>
<p>List of supporting tools/concepts (3-4 lines)</p>	<p>term frequency–inverse document frequency (TF-IDF) bag-of-words model (BoW) Convolutional neural network (CNN) bidirectional long short- term memory (LSTM) bidirectional gated recurrent unit (GRU)</p>
<p>What are the similar approaches with which the proposed approach is compared? (Refer experimental evaluation/result discussion) Explain each of these approach</p>	<p>Approach/method 1:Convolutional neural network (CNN) Discription: Approach/method 2: bidirectional long short- term memory (LSTM) Discription: Approach/method 3: bidirectional gated recurrent unit (GRU)</p>

<p>How the results of proposed approach are compared with other similar approaches? (Refer experimental evaluation/result discussion)</p>	<p>RVVC outperforms all other individual models when TF-IDF features are used with SMOTE balanced dataset and achieves an accuracy of 0.97</p>
<p>Advantages/merits of proposed solution in your view</p>	<p>Results suggest that balancing the data reduces the chances of models over-fitting which happens if the imbalanced dataset is used for training. Moreover, TF-IDF shows better classification accuracy for toxic comments than BoW as TF-IDF records the importance of a word contrary to BoW which simply counts the occurrence of a word</p>

<p>Disadvantages/limitations of proposed solution in your view. (Refer conclusion / result discussion / experimental evaluation)</p>	<p>Results indicate that models perform poorly on the imbalanced dataset while the balanced dataset tends to increase the classification accuracy</p>
<p>Your one page write-up about this paper</p>	
<p>Social media platforms and microblogging websites have gained accelerated popularity during the past few years. These platforms are used for expressing views and opinions about products, personalities, and events. Often during discussions and debates, fights take place on social media platforms which involves using rude, disrespectful, and hateful comments called toxic comments. The identification of toxic comments has been regarded as an essential element for social media platforms. This study introduces an ensemble approach, called regression vector voting classifier (RVVC), to identify the toxic comments on social media platforms. The ensemble merges the logistic regression and support vector classifier under soft voting criteria. Several experiments are performed on the imbalanced and balanced dataset to analyze the performance of the proposed approach. For data balance, the synthetic minority oversampling technique (SMOTE) is used on the imbalanced dataset. Furthermore, two feature extraction approaches are utilized to investigate their suitability such as term frequency-inverse document frequency</p>	

(TF-IDF) and bag-of-words (BoW). The performance of the proposed approach is compared with several machine learning classifiers using accuracy, precision, recall, and F1-score. Results suggest that RVVC outperform

Your findings: (possible alternate for the solution proposed)

- BERT Bidirectional Encoder Representations from Transformers
- CNN Convolutional Neural Network
- GRU Gated Recurrent Unit
- LSTM Long Short-Term Memory

TABLE-2 (ARTICLE 2)

Problem answered in this paper	To create an online interface where we would be able to identify the toxicity level in the given phrase or sentence and classify them into their order of toxicity
Detailed description about the problem	There have been sure turns of developments in this area which includes couple of models served through API. But the models still make errors and still fail to provide an accurate solution to the problem. In this paper we have widely discussed a set of models which is utilized for text classification. These models/methods have been widely used in various fields such as economics, medical and environmental studies.
Why that problem is chosen in this paper? Scope of the problem and solution (Refer Introduction)	Detecting Toxic comments has been a great challenge for the all the scholars in the field of research and development. This domain has drawn lot of interests not just because of the spread of hate but also people refraining people from participating in online forums which diversely affects for all the creators/content-providers to provide a relief to engage in a healthy public interaction which can be accessed by public without any hesitation.

History of the problem. (Refer Introduction)	Since the democratization of substance creation following the dispatch of web-based media stages, every single one of us has become content makers making and distributing our own
	substance, which thus has made a framework where the nature of distributed substance cannot, at this point be controlled. The effect of the most recent twenty years' innovation unrest is presently affecting organizations, political frameworks, family lives, society, and individuals
List of the related/similar problems (Refer Related work) – Describe each with proposed solutions	
Related problem 1 – Describe	
Paper in IEEE style	
What is the proposed solution in this paper for the problem chosen? (Refer Proposed work) (5-8 lines)	The proposed approach is called regression vector voting classifier (RVVC) and combines these models using soft voting criteria . The soft voting criteria ensure that the class with a high predicted probability by two classifiers will be considered as the final prediction.
Architecture of the proposed solution. (Refer proposed work) Diagram	<pre> graph TD A[/Kaggle Data set/] --> B[Pre Processing] subgraph B [Pre Processing] C[Remove whitespaces, digits, urls, stop words] D[Lemmatization] end B --> E[Classification Model Building] E --> F[Model Evaluation] F --> G[/Final Label toxic or not/] </pre>
Name of the approach as stated by the authors (if not, you try to give a name based on the concepts used)	Binary relevance method with Multinomial Naïve Bayes and Support vector classifier

<p>List of existing algorithms used by the authors to complete the proposed work. (1-2 lines for each algorithm)</p>	<p>logistic regression BR Method with Multinomial Naive Bayes classifiers BR Method with SVM classifier</p>
<p>List of datasets used. (Refer experimental evaluation/result discussion)</p>	<p>wikipedia_toxicity_subtypes</p>

<p>References/links to each of the dataset used in this paper (in IEEE style)</p>	<p>.https://www.tensorflow.org/datasets/catalog/wikipedia_toxicity_subtypes</p>
<p>List of equations that are very well applied in this problem domain</p>	<p>Equation 1 : Logistic regression</p>
<p>List of method(s)/metrics used to evaluate the proposed approach. (Refer experimental evaluation/result discussion)</p>	<p>BR Method with Multinomial Naive Bayes classifiers’ ‘BR Method with SVM classifier’</p>
<p>List of supporting tools/concepts (3-4 lines)</p>	<p>scikit-multilearn library</p>
<p>What are the similar approaches with which the proposed approach is compared? (Refer experimental evaluation/result discussion) Explain each of these approach</p>	<p>Approach/method 1: Convolutional Neural Networks(CNN) Approach/method 2: Recurrent Neural Network(RNN)</p>

How the results of proposed approach are compared with other similar approaches? (Refer <i>experimental evaluation/result discussion</i>)	The outcomes for the algorithms were as follows, if weif we compare both hamming losses, we could come to the conclusion that Naïve Bayes has a hamming loss of 3.6 and an accuracy of 87.6 whereas the hamming loss for SVM is 4.36 and the accuracy is 88.16. This gives us a brief insight to understand the optimal algorithm that can be utilized for ordering toxic comments
Advantages/merits of proposed solution in your view	Binary Relevance method with Multinomial Naive Bayes is an efficient algorithm that serves our purpose and has a hamming loss of 3.6 as compared to the hamming loss of SVM with a score of 4.36.
Disadvantages/limitations of proposed solution in your view. (Refer <i>conclusion / result</i>)	Results indicate that models perform poorly on the imbalanced dataset while the balanced dataset tends to increase the classification accuracy

<i>discussion / experimental evaluation</i>)	
Your one page write-up about this paper	

The increase in penetration of usage of internet services has increased exponentially in the past 4 months due to the ongoing pandemic, this has empowered an enormous number of dynamic new and old clients utilizing the web for different administrations ranging from academic, entertainment, industrial, monitoring and the emergence of a new trend in the corporate life i.e work-from-home. Due to this sudden emergence of the crowd using the web, there has been an ascent in the number of mischievous persons too. Now it is the primary task of every online platform provider to keep the conversations constructive and inclusive. The best example can be referred to, can be twitter, a web-based media stage where people share their views. This platform has already drawn a lot of flak because of the spread of hate speech, insults, threat, defamatory acts which becomes a challenge for many such online providers in regulating them. Thus, there is active research being conducted in the field of Toxic comment classification. Here we collate non-identical machine learning and other trivial techniques on the dataset and propose a model that outflanks all others and compares them one-on-one. We have undertaken the Kaggle dataset for the above reason which has been broadly used and one of the prime resources for scholars working in deducing the challenge of toxic comment classification. The results would help up to create an online interface where we would be able to identify the toxicity level in the given phrase or sentence and classify them into their order of toxicity

Your findings: (possible alternate for the solution proposed)

- CNN Convolutional Neural Network
- GRU Gated Recurrent Unit
- LSTM Long Short-Term Memory

5. Gaps Identified/Sub Problems

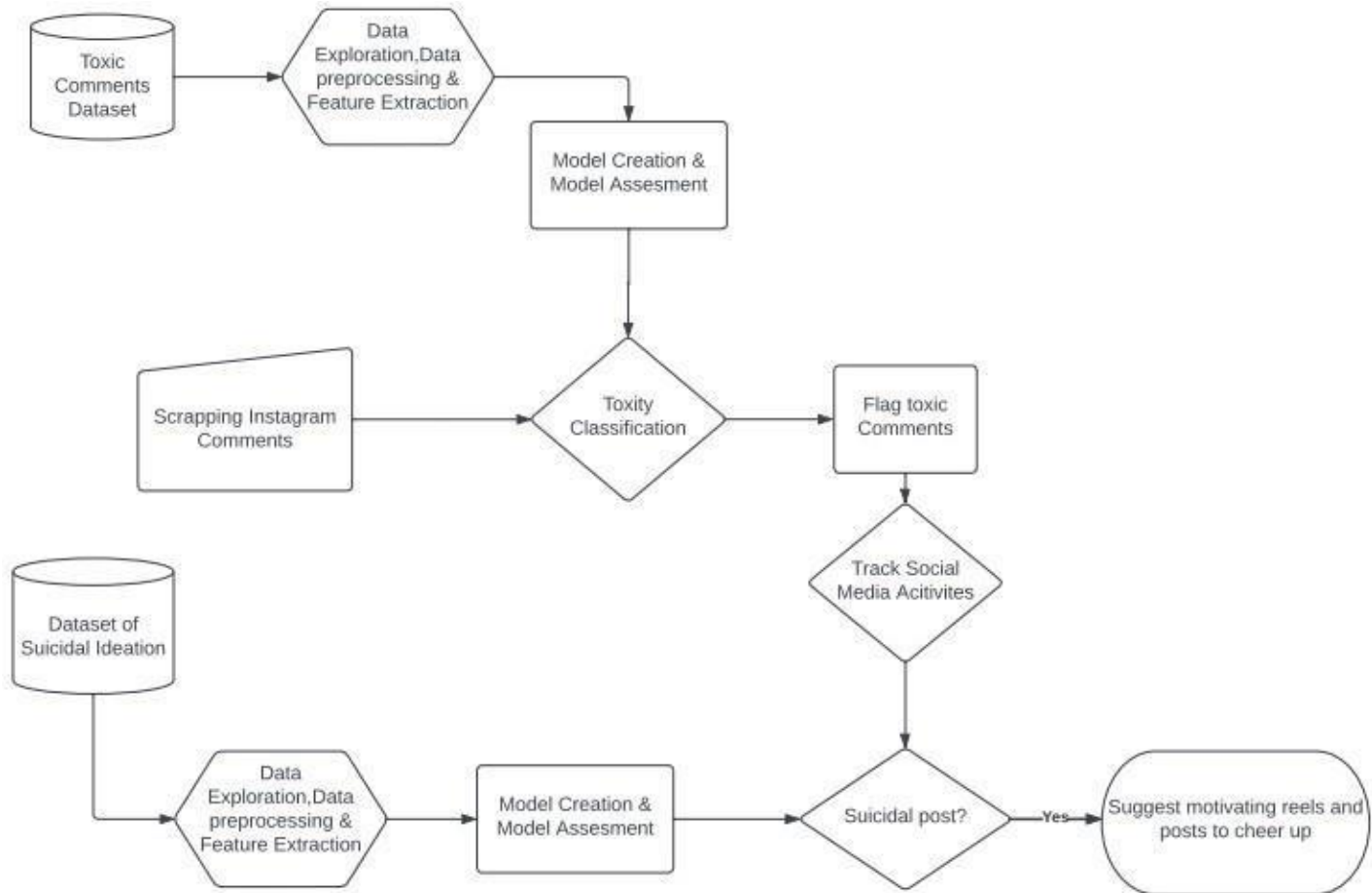
Here's an elaboration of the problem/ gap identified in previous research articles for Cyberbullying prevention by analyzing Social Media posts and comments using CNN and LSTM:

- **Multi-lingual Text Classification:** Cyberbullying can occur in multiple languages, and it is essential to identify such instances accurately. A system must be trained to classify text in different languages and detect any language-specific nuances of hate speech or cyberbullying. This can be challenging because some languages have unique writing systems or dialects that may require additional training data. For example, in India, the Hinglish dataset combines Hindi and English, which is a significant challenge for accurate classification. Therefore, a multilingual text classification system must account for these differences to ensure accurate detection and prevention of cyberbullying across languages.
- **Suicide Checking:** Cyberbullying can have severe consequences on the mental health of victims, including depression and suicidal tendencies. Therefore, it is crucial to identify instances of cyberbullying that may lead to self-harm or suicide ideation. A suicide checking system could analyze the tone of the text, identify any negative emotions or despair, and flag messages that suggest self-harm or suicide ideation. Such a system could alert the concerned authorities immediately, enabling them to take appropriate action to prevent harm to the victim.
- **Flagging and Blocking Culprits:** It is essential to identify and hold accountable those responsible for cyberbullying. A system could flag and block culprits who engage in abusive or harassing behavior online. The system would need to track and trace users who post harmful content and ensure that their profiles are flagged or blocked to prevent further harassment.
- **Tracking and Tracing of Users:** Cyberbullies often hide behind fake profiles or anonymous usernames. Therefore, it is challenging to identify such perpetrators and hold them accountable for their actions. A system that can track and trace the user who is doing the comment or posting the content would help law enforcement agencies to take action quickly against the perpetrator. The system could collect information about the user's IP address, geolocation data, and other identifying information to locate the individual responsible for the harassment.
- **Video Recommendation:** Videos are increasingly being used as a medium to spread hate speech and harass victims online. Identifying such videos and taking preemptive measures to prevent their circulation is essential for preventing cyberbullying. A video recommendation system could analyze the content of the videos to identify any content that could incite violence or hatred. Such a system could use CNN and LSTM models that are trained on a large dataset of videos to identify visual cues or patterns that suggest abuse or harassment. Based on the analysis, the system could recommend appropriate actions such as flagging the video or removing it from the platform.

In conclusion, preventing cyberbullying requires a comprehensive approach that involves NLP techniques, machine learning models such as CNN and LSTM, and advanced tracking and tracing mechanisms. A system that can accurately classify text in multiple languages, check for suicidal tendencies, flag and block perpetrators, track and trace users.

6. Proposed Model

5.1 Architecture



5.2 Explanation

A hybrid architecture of CNN-LSTM is proposed to classify toxic comments and check posts made by cyberbullying victim are suicidal or not. The work flow is as follows: □

- A neural network model using CNN-LSTM will be trained from toxic comments dataset taken from Kaggle Wikipedia's talk page edits dataset, reddit and twitter. □
- Data exploration , Data preprocessing and feature extraction will be done on the dataset used □
- User will input instagram ID and comments on posts will be extracted. □

- ✚ If the comments are toxic then users commenting will be flagged and if toxicity level is more than a 50% than activities of victim is tracked. □
- ✚ A neural network will be used to classify posts wheather they are suicidal or not. □
- ✚ If posts are classified as suicidal then motivational videos and reels will be suggested to victim.

7. Execution

7.1 Dataset Used

Dataset for detecting toxic comments was taken from kaggle “Dataset of comments from Wikipedia’s talk page edits”

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hat
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0	0	0	0	0	
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0	0	0	0	0	
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0	0	0	0	0	

Figure 1. Sample Rows of Original Dataset

It is a dataset of large number of Wikipedia comments which have been labeled by human raters for toxic behavior.

The types of toxicity are:

- ✧ toxic severe
- ✧ toxic obscene
- ✧ threat
- ✧ Insult
- ✧ Identity hate

The *Suicide and Depression Detection dataset* used for this project was obtained from Kaggle, which consists of posts from the social media platform Reddit.

	label	tweet
0	1	my life is meaningless i just want to end my l...
1	1	muttering i wanna die to myself daily for a fe...
2	1	work slave i really feel like my only purpose ...
3	1	i did something on the 2 of october i overdose...
4	1	i feel like no one cares i just want to die ma...

Figure 2. Sample Rows of Original Dataset

The dataset consists of 2 columns as seen in Figure 2 above, where tweet contains various tweets made by people and the other column contains 0 and 1. if tweet is suicidal it is labelled as 1 otherwise 0.

7.2 Metric Used

MEASURES USED FOR EVALUATION

⇒ Accuracy:

- ✧ This term tells us how many right classifications were made out of all the classifications.
- ✧ $\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$
 - ✧ True Positive (TP)
 - ✧ False Positive (FP)
 - ✧ True Negative (TN)
 - ✧ False Negative (FN)

After 2 epochs toxic comment detection model was giving a accuracy of 98.13%

In [67]:

```
model_info_1=model_1.fit(x_train,y_train, epochs=2, batch_size=32, validation_data=(x_val,
Epoch 1/2
3990/3990 [=====] - 633s 158ms/step - loss: 0.0699
- accuracy: 0.9030 - val_loss: 0.0511 - val_accuracy: 0.9939
Epoch 2/2
3990/3990 [=====] - 527s 132ms/step - loss: 0.0509
- accuracy: 0.9813 - val_loss: 0.0478 - val_accuracy: 0.9941
```

Whereas suicidal detection model was giving a accuracy of 91.2%

```
In [14]: print('Accuracy: %.3f' % clf.score(X_test, y_test))
Accuracy: 0.912
```

7.3 Modules Required

- ✧ Numpy
- ✧ Pandas
- ✧ Pickle
- ✧ Re
- ✧ Spacy
- ✧ Nltk
- ✧ Tensorflow
- ✧ Keras

7.4 Comparision Table

	Our project	Identification and Classification of Toxic Comment Using Machine Learning Methods	Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications
--	-------------	---	--

Datasets/corpora used	For toxic comment classification : Kaggle Wikipedia's talk page edits	wikipedia_toxicity_subtypes	<ul style="list-style-type: none"> sample of 500 titles of posts published in Reddit's Suicide Watch forum Twitter dataset for suicide risk assessment
data sources used	dataset For suicidal caption detection : Reddit Kaggle, Reddit	Wikipedia	<ul style="list-style-type: none"> Reddit Twitter
scope	A predictive web app was built to detect toxicity (accuracy=99.41) in user's instagram posts and detect suicidal intent (accuracy = 91%) in further instagram posts of victim. To further highlight the applicability of our webapp, we suggest motivational videos for victim.	To create an online interface where we would be able to identify the toxicity level in the given phrase or sentence and classify them into their order of toxicity	This paper aimed to describe an approach for the suicide risk assessment of Spanish-speaking users on social media. We aimed to explore behavioral, relational, and multimodal data extracted from multiple social platforms and develop machine learning models to detect users at risk.
important existing algorithms / approaches used	LSTM CNN Lemmatization Tokenization Stopwords Removal	Binary relevance method with Multinomial Naïve Bayes and Support vector classifier	BoW model trained with 1 to 5 grams deep learning model defined by a CNN architecture
supervised or unsupervised	supervised	supervised	Supervised
Accuracy	Toxicity model: 99.41 Suicidal model: 91%	Accuracy: 88.16%	Maximum Accuracy 94%

References

- [1] Elias Aboujaoude, Matthew W Savage, VladanStarcevic, and Wael O Salame. Cyberbullying: Review of an old problem gone viral. *Journal of adolescent health*, 57(1):10–18, 2015
- [2] How Much Data is Created on the Internet Each Day? <https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/>. Accessed: 2020-06-06
- [3] World Internet Users and 2020 Population Stats. <https://www.internetworldstats.com/stats.htm>. Accessed: 2020-06-06
- [4] Maeve Duggan. Online harassment. Pew Research Center, 2014
- [5] PinkeshBadjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017
- [6] Man jailed for 35 years in Thailand for insulting monarchy on Facebook. <https://www.theguardian.com/world/2017/jun/09/manjailed-for-35-years-inthailand-for-insulting-monarchy-on-facebook>. Accessed: 2020-06-06.
- [7] Mississippi teacher fired after racist Facebook post; black parent responds. <https://www.clarionledger.com/story/news/2017/09/20/mississippi-teacher-firedafter-racist-facebook-post/684264001/>. Accessed: 2020-06-06.
- [8] Ellery Wulczyn, NithumThain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391– 1399, 2017.
- [9] 85% Indian kids have experienced cyberbullying, highest in the world, finds new survey <https://theprint.in/india/85-indian-kids-haveexperienced-cyberbullyinghighest-in-the-world-finds-newsurvey/1074175/>

- [10] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Archives of suicide research*, vol. 14, no. 3, pp. 206–221, 2010.
- [11] Gliatto, M.F.; Rai, A.K. Evaluation and Treatment of Patients with Suicidal Ideation. *Am. Fam. Physician* 1999, 59, 1500. [PubMed]
- [12] Argyriou, A., & Tsolis, D. (2018). Detecting cyberbullying in social media using deep learning techniques. In 2018 IEEE International Conference on Cybersecurity and Privacy (Cybersecurity) (pp. 46-51). IEEE.
- [13] Gao, H., Liu, J., Zhang, S., & Chen, Q. (2019). A hybrid deep learning approach for cyberbullying detection on social media. *Multimedia Tools and Applications*, 78(18), 25933-25949.
- [14] Li, S., Sun, C., Huang, X., Tang, Y., & Wang, S. (2020). Cyberbullying detection in social media using a multi-task learning framework. *Information Fusion*, 60, 198208.
- [15] Mahmud, R., Shabut, A., Ramzan, N., Al-Turjman, F., & Shahzad, M. (2020). Deep learning-based cyberbullying detection in social media using text, image, and audio features. *Journal of Ambient Intelligence and Humanized Computing*, 11(6), 2287-2301.
- [16] Singh, S., Singh, S., & Bhatia, S. (2021). A comparative study of deep learning models for cyberbullying detection in social media. *Cognitive Systems Research*, 67, 102436.