PROTECT: Proactive Recognition of Offensive Texts, Images, Videos, and Memes Through AI

[1]Gurvinder Kaur, [2]Dhairvi Shah, [3]Utkarsha Kasar, [4]Anukriti Joshi,[5] Preeti Kale
[1] gurvinderkaur.mathura@mitwpu.edu.in [2] dhairvi.shah@mitwpu.edu.in [3] kasar.sunil@mitwpu.edu.in
[4] anukriti.joshi@mitwpu.edu.in [5] preeti.kale@mitwpu.edu.in
[1,2,3,4,5] Department of Computer Engineering & Technology, Dr. Vishwanath Karad MIT-World Peace University, Pune, India

ABSTRACT

The proliferation of images, videos, and textual content on social media as well as on other platforms (websites, blogs, articles) has engendered a multifaceted and varied online milieu. These modalities of content, frequently succinct and jocular, employ symbolism and humor to communicate concepts and ignite dialogues. Nevertheless, the simplicity of content generation also facilitates the dissemination of objectionable material, encompassing hate speech and stereotypes, thus requiring dependable detection methodologies. Natural Language Processing (NLP) furnishes a resolution by scrutinizing textual content to pinpoint inappropriate language. This investigation seeks to employ sophisticated machine learning algorithms to identify offensive components in images, videos, and textural data, taking into account visuals, emotions, and arrangements associated with detrimental content. Transcriptions of videos were incorporated for examination. The intricacy of content scrutiny transcends mere recognition of images or facial expressions, given that these mediums frequently employ irony, sarcasm, and cultural allusions. Consequently, contextually-aware tactics are imperative, entailing algorithms that discern customary patterns and incorporate sentiment analysis to evaluate overall emotional tone. In this study, various word embedding techniques: Word2Vec, BagOfWords, GloVe, FastText and Tf-Idf have been implemented after which the model was trained. The results obtained by using each technique were evaluated and compared.

Keywords: Word2vec, Bag of Words, Tf-Idf, GloVe, FastText, CNN, LSTM

1. INTRODUCTION

1.1 Motivation

The internet has grown into a vast resource of blogs, articles, and memes that swiftly transmit concepts and feelings by fusing text, humor, pictures, and videos. For the same, the internet has emerged as a productive field. These online resources are useful for social critique and communication, but there is a worrying subset of them that disseminates hate speech targeted at specific people or groups. Techniques for categorizing objectionable content need to be developed in order to promote a more cordial and friendly online community. The rise in hate speech An increasing number of negative content on the internet is a cause for concern. Research indicates a noteworthy surge in hate speech and abuse on the internet, often disguising itself as memes, articles, and blogs. These internet sources have the power to spread misinformation, inspire violence, and create a dangerous environment for those who are marginalized, thus it is crucial to pay attention to them.

The complexity of these resources lies in their multimodal composition. The influence depends on the interaction between text and image/video. Focusing solely on one element neglects the full story conveyed. Utilizing classification methods that consider both text and image/video is crucial.

- the visual and the textual
- to effectively pinpoint hateful content.

Understanding nuances and context is crucial in identifying hateful content which often use sarcasm, irony, or cultural references. Differentiating authentic hate speech from satire or social commentary requires grasping the subtleties in language and visual content. Classification methods should go beyond simple keyword matching and take into account the complex interplay between meaning and purpose.

Navigating ambiguity: Not all internet contents targeting a particular group are inherently hateful. They can be instrumental in raising awareness about social issues or reclaiming offensive language. Classification models must exhibit sophistication in discerning between these categories and genuinely malicious content.

Advancing safety and civility: Efficiently categorizing hateful internet resources empowers online platforms to take proactive measures. This may involve content moderation, flagging objectionable material, or guiding users towards resources that advocate online safety and mutual respect. Comprehending societal trends: Hateful internet resources can act as a mirror reflecting societal anxieties and biases. Analyzing the content and dissemination of these internet resources offers valuable insights for researchers and policymakers striving to combat hate speech and foster social harmony.

Cultivating a more inclusive online community: Eliminating hateful internet content from the web nurtures a welcoming online sphere where individuals feel at ease expressing themselves. This holds particular significance for marginalized groups that often bear the brunt of online animosity. This work may contribute in the near future to the wider fields of AI and NLP. Progressing towards robust models for classifying hateful internet resources necessitates advancements in fields such as multimodal learning, sentiment analysis, and sarcasm detection. These advancements can subsequently be harnessed for diverse applications like enhancing online safety mechanisms or refining conversational chatbots.

1.2. Objectives:

The objectives of this study are in a multi-stage process involving:

1.2.1. Train the model on Labeled dataset
1.2.2. Use the trained model to classify unseen data
1.2.3. Paraphrase the offensive content

This objective highlights the main goal of the research, which is to create a system that can effectively identify and categorize offensive content specifically within the context of Indian memes. It details how to employ deep learning architectures to create potent classification models, transfer learning to gain from models that were previously trained, and use NLP (natural language processing) for text analysis This objective implicitly acknowledges the real-world problem that this research aims to contribute to solving. Hate speech can have significant negative consequences, and the ability to automatically classify it can be valuable for social media platforms and other online communities.

2. LITERATURE SURVEY

Study	Dataset	Techniques Applied	Application	Research Gap
[1]	MMHS150k dataset	Deep learning, machine learning, statistical biases, Graph Neural Networks	Hateful memes detection	Neglect of multimodal techniques and demographic considerations in prior studies
[2]	Amazon camera product reviews	Data mining, opinion mining, Naive Bayes method, sentiment analysis	Gauging public sentiment, public health monitoring	Real-Time Sentiment Monitoring
[3]	Hateful Memes Challenge dataset	Pre-trained models (ResNet-50, CLIP), PyTorch CUDA, ensemble learning	Abusive speech detection in online memes	Contextual Understanding, Multimodal Fusion Techniques, Ethical Considerations, Real-time Processing
[4]	CMU-MOSI: 2199 videos, individual commentary. CMU-MOSEI: 23453 clips, 5000 videos.	Natural language processing, preprocessing techniques (median filtering, MFCC), TF-IDF, word embedding, LSTM model	Sentiment analysis of social media comments	Enhancing multimodal sentiment with NNs.
[5]	OCR combined with Naive Bayes algorithm	OCR, Neural Networks, K-Nearest Neighbor algorithms, n-gram tokenization	Sentiment analysis of memes	Limitations of OCR Tesseract Engine, potential methods for enhanced accuracy

[6]	IMDB, IEMOCAP, Twitter datasets	Keyword Spotting System, Singular Value Decomposition, Maximum Entropy classification algorithm	Sentiment analysis of YouTube videos	Addressing challenges in automatic voice recognition, improving accuracy
[7]	Twitter dataset	Lexicon-based methods, machine learning methods (BERT, BiLSTM, RNN, LSTM), statistical analysis	Sentiment analysis on social media text	Improving sentiment analysis accuracy, understanding dynamics of social media sentiment
[8]	Online conversations dataset	Convolutional Neural Networks, ensemble learning, support vector machines, lexicon-based techniques	Detecting sarcasm in online conversations	Capturing nuances of sarcasm effectively, mitigating adverse effects on social platforms
[9]	MultiOFF dataset	Multimodal strategy, ensemble learning, class imbalance mitigation methods	Offensive content detection in multimodal memes	Enhancing offensive content identification, utilizing multimodal techniques
[10]	Dataset of 10,000 memes	Multimodal reasoning, Cohen's kappa score analysis	Hate speech detection	Addressing limitations of current state-of-the-art models, advancing multimodal research
[12]	Twitter dataset	Naive Bayes, Support Vector Machine, Decision Tree, K-nearest neighbor classifiers	Sentiment analysis of Twitter posts	Improving sentiment analysis accuracy, classifying effectiveness for Twitter sentiment analysis

Table 1: Overall Summary of the Literature Review

The presented research in Table 1, encompasses a wide array of sentiment analysis applications spanning various datasets and domains. They utilize a variety of methods, including deep learning, machine learning, natural language processing, and ensemble learning, to tackle specific issues like identifying hate speech in memes, assessing public sentiment from product reviews, and examining sentiment in social media text and online conversations. Notable gaps in research include the oversight of multimodal approaches and demographic factors, constraints in existing sentiment analysis models, and the necessity for real-time processing and ethical considerations. In general, these investigations contribute to the progression of sentiment analysis techniques, improvement of model efficacy, and addressing of emerging challenges in comprehending and deciphering sentiment across diverse media and scenarios.

3. PROTECT Algorithm

The PROTECT algorithm delineates the comprehensive design flow implemented throughout this project. It encompasses every stage, from data collection and preprocessing to training and paraphrasing. The detailed workflow is outlined as follows:

PROTECT ALGORITHM:
Step 1: Dataset Acquisition & Normalization
Step 2: Data Preprocessing
Step 3: Word Cloud
Step 4: Word Embeddings
Step 5: Model Training
Step 6: Classifying Unlabeled Data
Step 7: Paraphrasing

The PROTECT algorithm comprises a systematic sequence of steps designed to optimize data processing and analysis. Initially, Dataset Normalization ensures that all data is converted into a consistent format, facilitating seamless integration and comparison. Subsequently, Data Preprocessing is performed to enhance data quality through cleaning, transformation, and augmentation. Following this, a Word Cloud is generated to provide a visual representation of word frequency and significance within the dataset. Next, Word Embeddings are created to capture semantic relationships between words, enabling more nuanced data interpretation. The Model Training phase involves using machine learning techniques to build predictive models based on the preprocessed data. Once trained, these models are employed for Classifying Unlabeled Data, extending their utility to new, unseen instances. Finally, the Paraphrasing step refines the output by rephrasing and summarizing the classified data, ensuring clarity and coherence in the results.

4. METHODOLOGY

This study is dedicated to developing a robust and dependable model capable of accurately identifying offensive content, particularly within the Indian context. Our focus encompasses not only memes but also images, videos, text articles, and blogs. In our study, we have chosen a strategy that combines transfer learning techniques with cutting-edge methods from NLP to accomplish this aim. This strategic combination is intended to enhance both the precision and efficiency of the classification process. The ultimate objective is to offer a holistic solution that can effectively detect and classify hateful content across various media types, thereby making a valuable contribution to broader initiatives aimed at reducing online hate speech and fostering a more secure digital space.

The different steps that were taken in this study and their detailed version can be seen in Figure 1

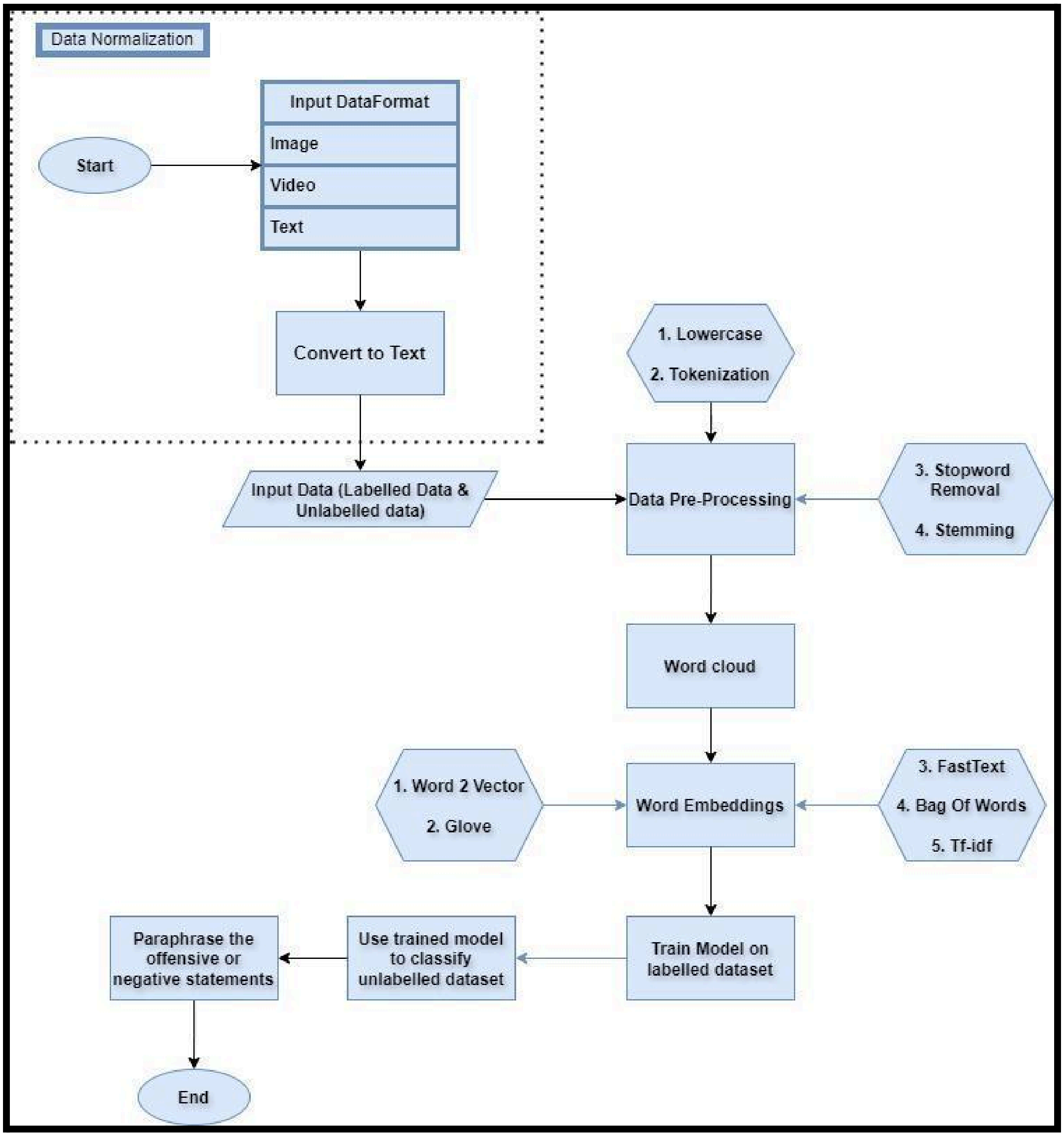


Figure 1: PROTECT Work Flow

4.1 Dataset Acquisition and Normalization

We intend to utilize the dataset openly available at [14], which contains annotations related to memes, with a specific focus on Indian cultural nuances. This dataset holds immense importance for our team as it serves as a valuable resource for not only creating but also evaluating the performance of our hate speech identification algorithm. Our goal is to improve the precision and effectiveness of our classification system by making use of this dataset, especially in the context of detecting hate speech within the Indian cultural framework.

In this study, our dataset comprises memes, images, video transcriptions, and text files sourced from diverse articles and blogs.

- Notably, text or captions extracted from images are meticulously recorded
- Video content is transcribed frame by frame
- Text files are directly incorporated into the dataset

This meticulous aggregation process culminates in the conversion of all data modalities into textual representations, facilitating subsequent natural language processing (NLP) tasks. This can be represented in figure 2 given below. Leveraging this dataset will enable us to strengthen our capabilities in tackling hate speech more effectively and efficiently in the Indian context.

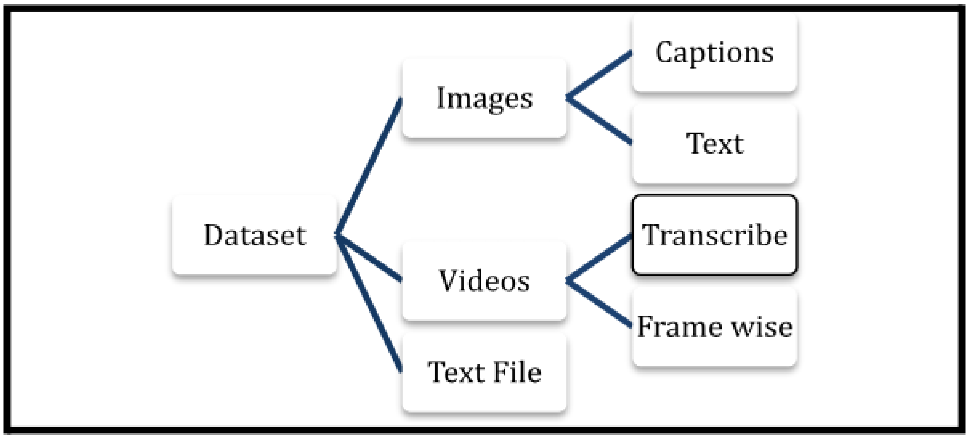


Fig 2 : Dataset Included

After processing images, videos, and textual datasets, all data are converted to a text format to facilitate more efficient utilization. Consequently, the data are normalized to a standardized textual format.

4.2 Data Preprocessing

Lowercase: Converting all text to lowercase ensures consistency in the data and avoids treating "Cat" and "cat" as different words.

Tokenization: By doing this, the text is divided into understandable chunks, usually words or characters.. This allows for further analysis like counting word frequencies or building vocabulary for models.

Stopword Removal: Stopwords are very common words that carry little meaning on their own, such as "the", "a", "an", or "is" By concentrating on content-specific terms that are more significant in interpreting the tone or meaning of the text, removing these words might increase the effectiveness of NLP jobs.

Stemming: This technique reduces words to their base form. For instance, "running" would be transformed to "run" using stemming. Stemming typically uses a set of rules to chop off suffixes, and while it's efficient, it can sometimes create grammatically incorrect words (e.g., "stemming" becomes "stemm").

4.3 Word Cloud

Word clouds provide an instant visual representation of the most common terms in the corpus by condensing textual data into an eye-catching shape in which word size reflects frequency. Word clouds give a brief summary of the main ideas and subjects covered in the book by highlighting commonly occurring words and underlining less common ones. This makes understanding and analysis easier and faster. Figure 3 and 4 depicts the word cloud for labeled and unlabeled dataset respectively.

For Labeled Dataset:

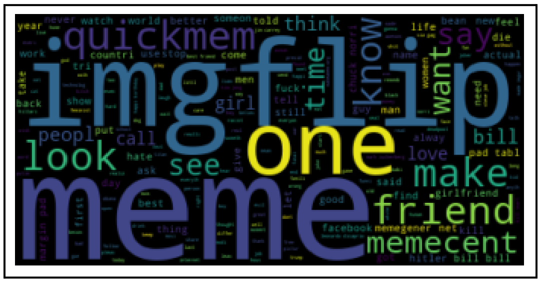


Fig 3: Word Cloud for Labeled Dataset

For Unlabeled Dataset:

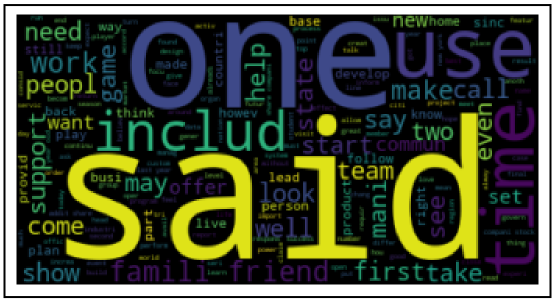


Fig 4: Word Cloud for Unlabeled Dataset

4.4 Word Embedding:

Word2Vec, a pivotal tool in the realm of Natural Language Processing (NLP), excels at encoding words by analyzing their contextual surroundings within an extensive textual dataset. This technique functions using two primary frameworks:

- Skip-gram, which anticipates neighboring words based on a specified word
- CBOW (Continuous Bag-of-Words), which predicts a specified word using its neighboring words.

Word2Vec's main benefits are how easy it is to use and how well it captures word semantic connections.Nevertheless, it encounters challenges with infrequent words and overlooks the internal structure of a word.

GloVe, known as the champion of Global Vectors, adopts a statistical methodology by examining the co-occurrence frequency of words in a corpus to derive word vectors. This strategy enables GloVe to excel in handling unfamiliar terms, i.e., words absent in the training data. Furthermore, it can potentially grasp certain facets of word similarity based on context. However, training GloVe on extensive datasets may incur high computational costs, and it might not capture intricate semantic connections as effectively as Word2Vec.

FastText emerges as the frontrunner in addressing uncommon and out-of-vocabulary (OOV) words. It extends the groundwork laid by Word2Vec but integrates subword details (n-grams) into the vector representation. By deconstructing words into smaller character sequences and learning vector representations for these n-grams, FastText adeptly manages unfamiliar words. This proves especially advantageous for languages characterized by intricate morphology, where word semantics can vary based on prefixes or suffixes. Nevertheless, the training process for FastText may be slower due to processing n-grams, and it might not depict robust semantic relationships for common words compared to Word2Vec.

Bag of Words (BoW) provides a numerical representation of the instance count of an expression in a text . It ignores word order and concentrates only on word frequency. BoW is a straightforward yet efficient method of text analysis, although it is devoid of context.

TF-IDF determines the importance of a word in a document by looking at how regularly it appears in a corpus. It is divided into two sections:

- inverse document frequency
- term frequency

To assist find important phrases, TF-IDF identifies words that appear often in a manuscript but are uncommon in the corpus.

4.5 Model Training

Machine learning (ML) and Advanced Machine Learning algorithmic techniques were employed to train on a labeled dataset, aiming to extract meaningful patterns and relationships from the data. This process involved utilizing various algorithms and methodologies to build predictive models capable of generalizing from the training data to make accurate predictions on unseen data.

1. Basic Artificial Neural Network

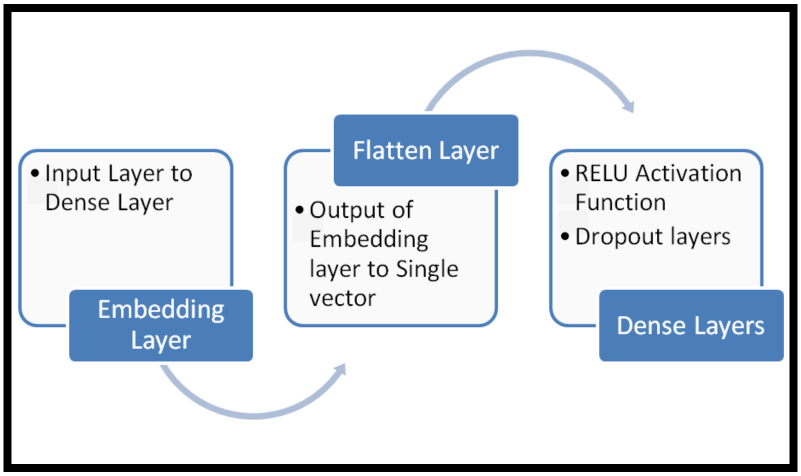


Fig 5: ANN Architecture

It had three main parts as depicted in the above figure 5 and ReLu was the activation function adopted for this particular strategy.

2. Simple LSTM

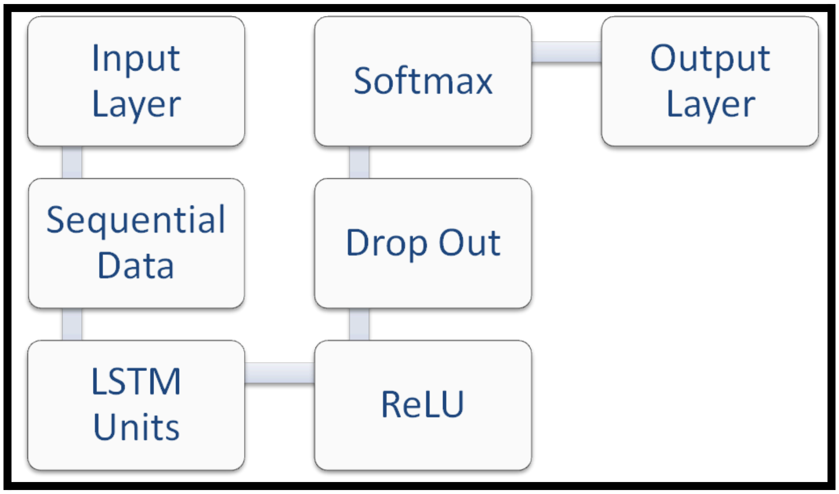


Fig 6: LSTM Architecture

LSTM units are feeded the sequential data for its training and then dense layers are the same as the simple ANN structure. As shown in figure 6 Softmax is also added to do multiple class classification.

3. Hyperparamter Tuned LSTM

Here the same Simple LSTM is applied but with fine tuned parameters. The parameters used for this method are as shown in figure 7:

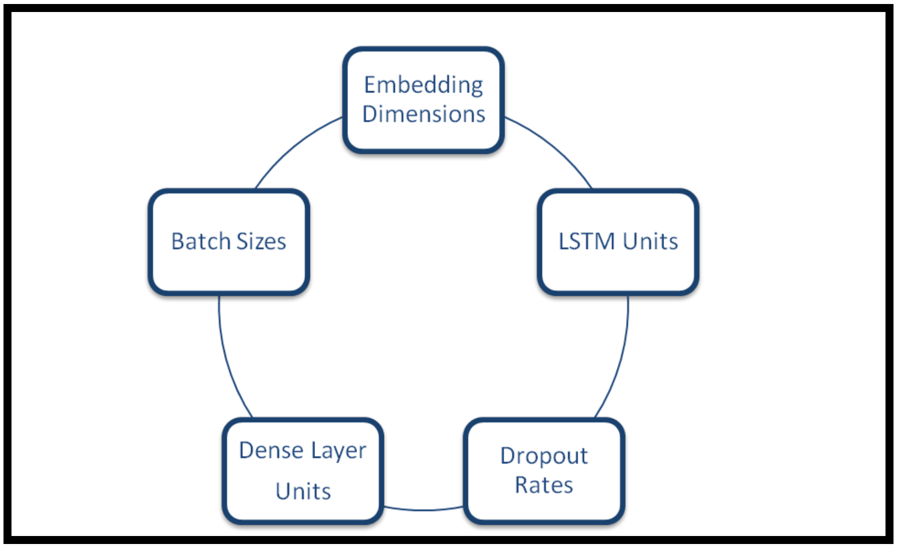


Fig 7: Hyperparameter Tuned LSTM

The parameters were :

1. Embedding_dims = [50,100,150]
2. Lstm_units = [64,128]
3. Dropout_rates = [0.2,0.5]
4. Dense_units = [32,64]
5. Batch_sizes = [32,64]

And the total number of epochs for the method was set to 5.

4. CNN along with LSTM embedded within

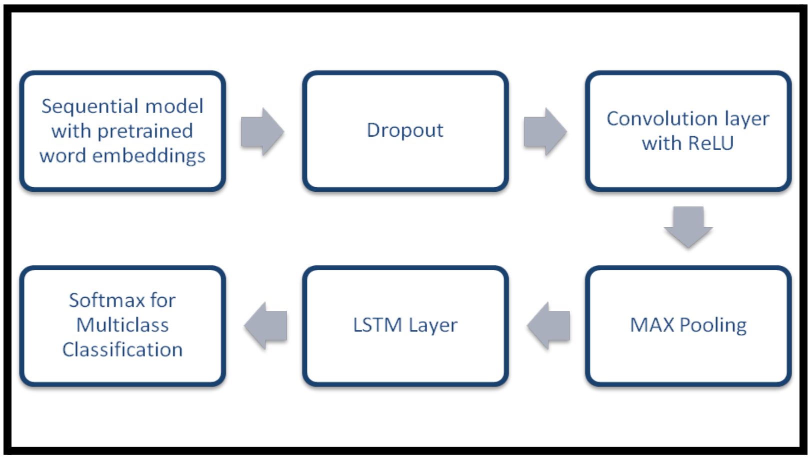

Sequential model with pretrained word embeddings
Dropout
Convolution layer with ReLU
MAX Pooling
LSTM Layer
Softmax for Multiclass Classification

Fig 8: CNN with LSTM

Dropout Layer at an early stage as shown in figure 8 proves to be quite better in terms of providing results along with the Convolution layers added in this method.

4.6 Classifying Unlabelled Data

The algorithm's stored model is used to categorize unlabeled data after the training phase. The same word embedding strategies and NLP preprocessing methods that were used in the training phase must be applied in this procedure. Lower casing, Tokenization, stemming, and stopword removal are just a few of the preprocessing procedures that turn unstructured text input into a format that may be used for computational analysis. Word embeddings are also used to capture contextual information and semantic linkages in the text by representing words as tightly packed vectors in a space with high dimensions. The model maintains coherence and consistency in interpreting and categorization of unlabelled data by using the same word embeddings and preprocessing methods as in the training phase. This makes textual content analysis more accurate and meaningful.

4.7 Paraphrasing

Following the classification of offensive or negative statements, a paraphrasing process is initiated to transform the tone of these statements into a more positive demeanor, while endeavoring to maintain the original meaning intact. This involves a meticulous linguistic exercise aimed at rephrasing the content in a manner that imbues it with positivity and optimism without deviating from the intended message. By carefully selecting synonymous expressions, restructuring sentences, and infusing affirmative language, the paraphrased statements are crafted to convey a sense of encouragement, support, or affirmation. This delicate balance between preserving the essence of the original statement and transmuting its tone underscores the nuanced nature of the paraphrasing endeavor, which seeks to enhance the overall positivity and constructiveness of the discourse while respecting the integrity of the underlying content.

5. RESULTS AND DISCUSSIONS

5.1 Trained Model

In order to identify the best-performing model, we evaluated a variety of word embedding methods, such as Word2Vec, TF-IDF, Bag of Words, GloVe, and FastText. Following this initial stage, an CNN architecture combined with a series of LSTM units was used to train each model. When evaluating the effectiveness of these models, our primary focus was on accuracy as a key metric for performance assessment. Accuracy, within this context, denotes the model's capacity to accurately categorize instances within the dataset.

Word Embedding	Accuracy	Val_Accuracy
FastText	93.07	70.67
GloVe	99.68	78.72
Bag Of Words	94.02	79.09

Tf-idf	94.145	75.9
Word2Vec	96.02	81.01

Table 2: Result of all the methods used in this study

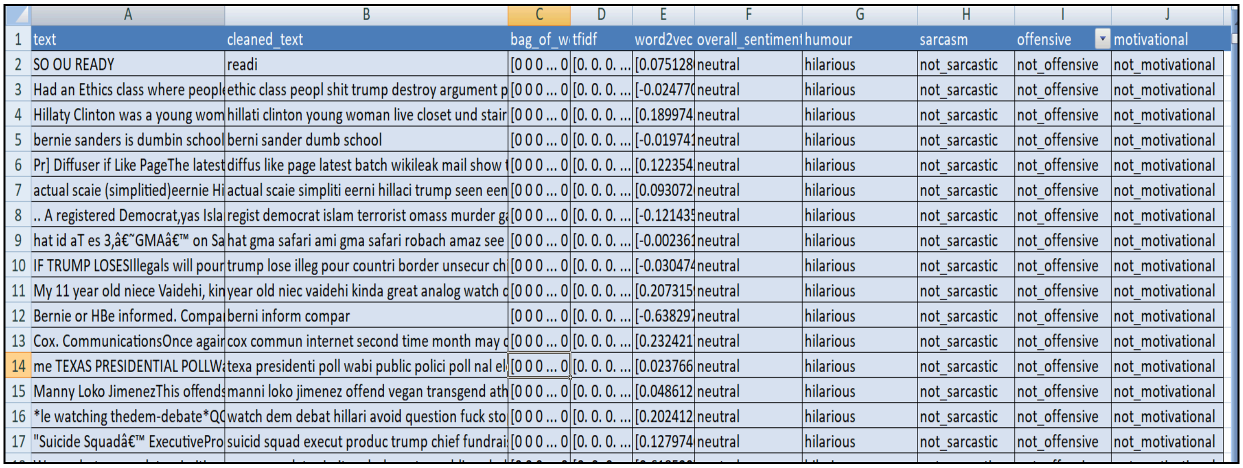
From table 2 we can infer that during the examination of diverse word embedding approaches, GloVe shows the highest level of precision at 99.68%, exceeding alternatives such as FastText, Word2Vec, Bag of Words, and Tf-idf. Nonetheless, FastText attains a respectable precision of 93.07%, whereas Word2Vec closely trails at 96.02%. The preciseness rankings of Bag of Words and Tf-idf systems are almost equal, with Bag of Words slightly edging out Tf-idf. Upon analyzing validation accuracy, comparable patterns arise, with GloVe retaining the highest efficacy, albeit with a lesser relative enhancement over other methodologies in comparison to the training accuracy. In general, GloVe excels due to its outstanding precision, followed by Word2Vec and FastText, while Bag of Words and Tf-idf offer viable options for word embedding in tasks related to natural language processing.

5.2 Classifying the Unlabeled Dataset

The subsequent outcome shown in figure 8, yielded a fully labeled dataset, encompassing textual content sourced from diverse mediums, including images, videos, articles, websites, and blogs. Each entry in this dataset has been meticulously categorized based on its thematic essence, specifically classified into categories denoting whether the content is offensive, humorous, negative, or neutral in nature.

This annotated dataset stands as a comprehensive repository, reflecting a spectrum of textual expressions encountered across various digital platforms. It encompasses a rich array of linguistic forms, ranging from informal colloquialisms to structured prose, thereby presenting a nuanced portrayal of language usage within contemporary digital discourse.

The process of labeling involved subjecting each piece of textual content to rigorous scrutiny, guided by predefined criteria delineating the distinct thematic attributes under consideration. Through this meticulous annotation process, annotators assigned appropriate labels to each text segment, capturing its inherent tone and semantic orientation with precision.

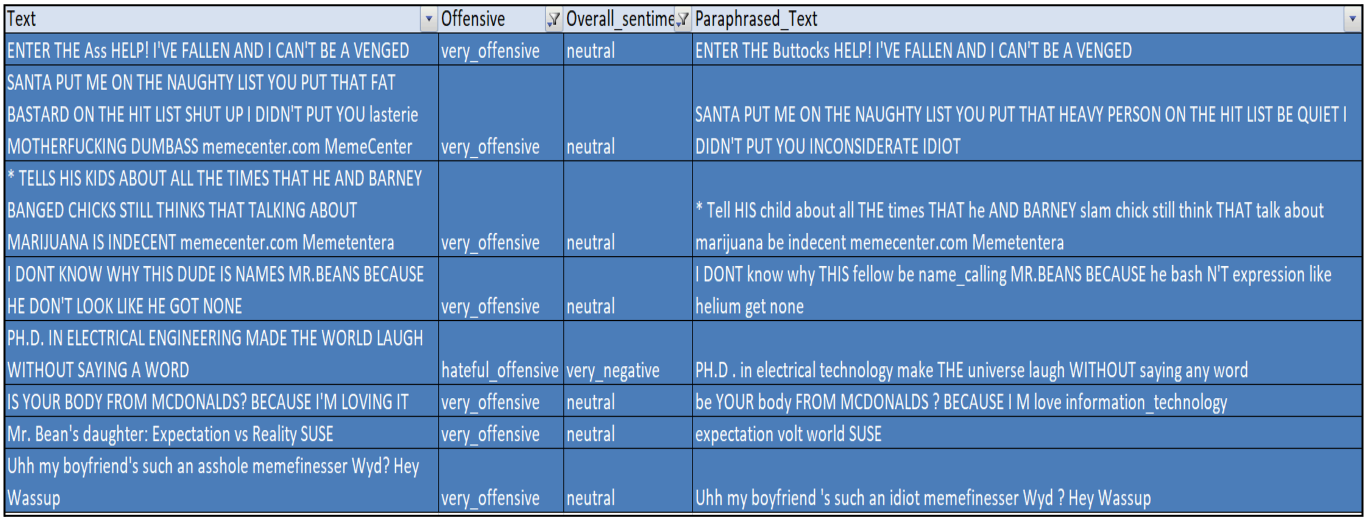


	A	B	C	D	E	F	G	H	I	J
1	text	cleaned_text	bag_of_w	tfidf	word2vec	overall_sentiment	humour	sarcasm	offensive	motivational
2	SO OU READY	readi	[0 0 0 ... 0	[0. 0. 0. ...	[0.075128	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
3	Had an Ethics class where people	ethic class peopl shit trump destroy argument p	[0 0 0 ... 0	[0. 0. 0. ...	[-0.024770	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
4	Hillaty Clinton was a young wom	hillati clinton young woman live closet und stair	[0 0 0 ... 0	[0. 0. 0. ...	[0.189974	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
5	bernie sanders is dumbin school	berni sander dumb school	[0 0 0 ... 0	[0. 0. 0. ...	[-0.019741	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
6	Pr] Diffuser if Like PageThe latest	diffus like page latest batch wikileak mail show t	[0 0 0 ... 0	[0. 0. 0. ...	[0.122354	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
7	actual scaie (simplitied)eernie Hi	actual scaie simpliti eerni hillaci trump seen een	[0 0 0 ... 0	[0. 0. 0. ...	[0.093072	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
8	.. A registered Democrat,yas Isla	regist democrat islam terrorist omass murder ga	[0 0 0 ... 0	[0. 0. 0. ...	[-0.121435	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
9	hat id aT es 3,â€˜GMAâ€™ on Sa	hat gma safari ami gma safari robach amaz see	[0 0 0 ... 0	[0. 0. 0. ...	[-0.002361	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
10	IF TRUMP LOSESIllegals will pour	trump lose illeg pour countri border unsecur chi	[0 0 0 ... 0	[0. 0. 0. ...	[-0.030474	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
11	My 11 year old niece Vaidehi, kin	year old niec vaidehi kinda great analog watch c	[0 0 0 ... 0	[0. 0. 0. ...	[0.207315	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
12	Bernie or HBe informed. Compa	berni inform compar	[0 0 0 ... 0	[0. 0. 0. ...	[-0.638297	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
13	Cox. CommunicationsOnce agair	cox commun internet second time month may c	[0 0 0 ... 0	[0. 0. 0. ...	[0.232421	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
14	me TEXAS PRESIDENTIAL POLLWa	texa presidenti poll wabi public polici poll nal el	[0 0 0 ... 0	[0. 0. 0. ...	[0.023766	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
15	Manny Loko JimenezThis offends	manni loko jimenez offend vegan transgend ath	[0 0 0 ... 0	[0. 0. 0. ...	[0.048612	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
16	*le watching thedem-debate*QC	watch dem debat hillari avoid question fuck sto	[0 0 0 ... 0	[0. 0. 0. ...	[0.202412	neutral	hilarious	not_sarcastic	not_offensive	not_motivational
17	"Suicide Squadâ€™ ExecutivePro	suicid squad execut produc trump chief fundrai	[0 0 0 ... 0	[0. 0. 0. ...	[0.127974	neutral	hilarious	not_sarcastic	not_offensive	not_motivational

Fig 8: Classifying Unlabeled Data

5.3 Paraphrasing

Upon successful classification of both the labeled and unlabeled datasets, a subsequent phase of the project involved the generation of paraphrases for the textual content. Notably, within both datasets, segments bearing labels denoting negative sentiments, including "Slight," "Very_offensive," "Hateful_offensive," "Negative," and "Very_negative," underwent a transformation aimed at imbuing them with a more positive tone, albeit with alterations in language structure and expression. This paraphrasing process entailed the careful rephrasing of text segments characterized by negative sentiments, endeavoring to infuse them with a more optimistic outlook while preserving the essence of the original content and it can be seen in Figure 9.



Text	Offensive	Overall_sentime	Paraphrased_Text
ENTER THE Ass HELP! I'VE FALLEN AND I CAN'T BE A VENGED	very_offensive	neutral	ENTER THE Buttocks HELP! I'VE FALLEN AND I CAN'T BE A VENGED
SANTA PUT ME ON THE NAUGHTY LIST YOU PUT THAT FAT BASTARD ON THE HIT LIST SHUT UP I DIDN'T PUT YOU lasterie MOTHERFUCKING DUMBASS memecenter.com MemeCenter	very_offensive	neutral	SANTA PUT ME ON THE NAUGHTY LIST YOU PUT THAT HEAVY PERSON ON THE HIT LIST BE QUIET I DIDN'T PUT YOU INCONSIDERATE IDIOT
* TELLS HIS KIDS ABOUT ALL THE TIMES THAT HE AND BARNEY BANGED CHICKS STILL THINKS THAT TALKING ABOUT MARIJUANA IS INDECENT memecenter.com Memetentera	very_offensive	neutral	* Tell HIS child about all THE times THAT he AND BARNEY slam chick still think THAT talk about marijuana be indecent memecenter.com Memetentera
I DONT KNOW WHY THIS DUDE IS NAMES MR.BEANS BECAUSE HE DON'T LOOK LIKE HE GOT NONE	very_offensive	neutral	I DONT know why THIS fellow be name_calling MR.BEANS BECAUSE he bash N'T expression like helium get none
PH.D. IN ELECTRICAL ENGINEERING MADE THE WORLD LAUGH WITHOUT SAYING A WORD	hateful_offensive	very_negative	PH.D . in electrical technology make THE universe laugh WITHOUT saying any word
IS YOUR BODY FROM MCDONALDS? BECAUSE I'M LOVING IT	very_offensive	neutral	be YOUR body FROM MCDONALDS ? BECAUSE I M love information_technology
Mr. Bean's daughter: Expectation vs Reality SUSE	very_offensive	neutral	expectation volt world SUSE
Uhh my boyfriend's such an asshole memefinesser Wyd? Hey Wassup	very_offensive	neutral	Uhh my boyfriend 's such an idiot memefinesser Wyd ? Hey Wassup

Fig 9: Paraphrased original data

Through strategic linguistic adjustments and semantic reinterpretations, the transformed text segments were crafted to convey a sense of positivity and optimism, thereby mitigating the underlying negativity inherent in the original expressions. By systematically altering the language and tone of negatively labeled text segments, the paraphrasing endeavor aimed to foster a more balanced and uplifting narrative within both the labeled and unlabeled datasets. This approach not only contributes to the enhancement of textual diversity and sentiment balance within the datasets but also underscores the potential of natural language processing techniques in fostering constructive discourse and promoting positivity within digital communication channels.

Figures 10 and 11 below illustrate the categorization of the text data based on the labels "Offensive" and "Overall Sentiments," respectively. The labels "Slight," "Very_offensive," "Hateful_offensive," "Negative," and "Very_negative," are paraphrased as shown in figure 9.

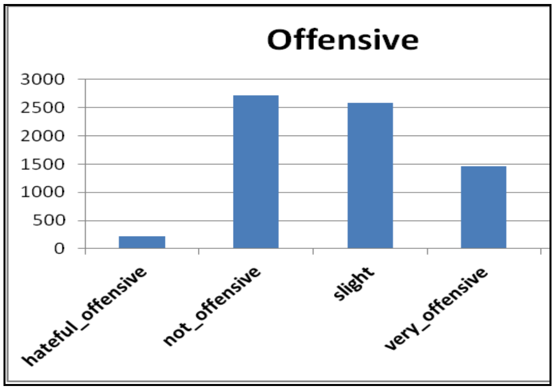


Fig 10: Offensive Labels

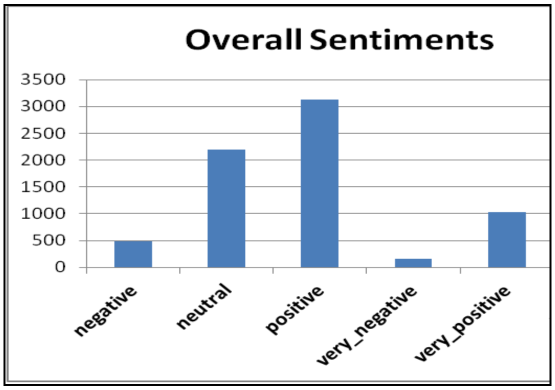


Fig 11: Overall Sentiment Labels

6. CONCLUSION and FUTURE SCOPE

This research project proposes a novel system aimed at identifying and curtailing the dissemination of offensive content pervasive across digital media platforms. With the use of deep learning and NLP algorithms, the algorithm analyzes text and visual content in a wide range of formats, including blogs, articles, videos, memes, and photos Through this synergistic approach, it detects harmful content that may evade traditional moderation methods, recognizing hateful symbols, discriminatory visual cues, offensive phrases, and sarcasm with malicious intent. The envisioned system goes beyond mere identification, striving to categorize offensive content based on severity, cultural context, and potential harm. This categorization empowers platforms to implement targeted interventions, fostering a more positive and inclusive online environment where users can engage freely without fear of encountering offensive material.

However, challenges such as algorithmic bias and subjective interpretations of humor versus offense must be addressed. Mitigation strategies include ensuring diverse and representative training data, ongoing monitoring, and incorporating user feedback mechanisms for borderline cases. Despite these challenges, the proposed system represents a significant stride towards responsible online moderation, leveraging advanced technologies to promote respectful discourse and combat online harassment and discrimination.

Future Scope:

Future iterations of the recommended technique can be developed in a variety of ways. Continued research into multimodal learning techniques can enhance the model's ability to analyze the intricate interplay between text and imagery within diverse digital media. Additionally, advancements in sentiment analysis and sarcasm detection can further improve the model's contextual understanding, enabling it to more accurately discern the intent behind nuanced expressions. Furthermore, investigating the possibilities of explainable AI approaches might improve accountability and transparency in the algorithm's decision-making process, building user and stakeholder trust. Incorporating this strategy into other forms of digital content, such text-based postings and comments, can boost the suggested method's efficacy in curbing cyber bullying and fostering a friendlier and more respectful virtual community.

REFERENCES

1. A. Sethi, U. Kuchhal, R. Katarya, and A. Anjum, "Study of Various Techniques for the Classification of Hateful Memes," 2021 6th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), pp. 675-678, Aug. 2021, doi: 10.1109/RTEICT52883.2021.978-1-6654-3559-8/21/$31.00 ©2021 IEEE.
2. S. R. Santhosh, V. S. Vignesh, R. E. Rithish, and M. P. D. Mahendhiran, "Sentimental Analysis on Amazon Camera Reviews using Naive Bayes Algorithm," 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), pp. 1542-1545, Feb. 2023, doi: 10.1109/ICSCSS57650.2023.10169302.
3. G. Macrayo, W. Caŝino Jr, J. Dalangin, J.G. Gahoy, A.C. Reyes, C. Vitto, M. Abisado, S.L. Huyo-a, and G.A. Sampedro, "Please Be Nice: A Deep Learning Based Approach to Content Moderation of Internet Memes," 2023 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1-6, Jan. 2023, doi: 10.1109/ICEIC57457.2023.10049865.
4. B. Sravani, P. Mohan, A.H.A. Hussein, G.R. Kumar, and P. Umaeswari, "Multimodal Sentimental Classification using Long-Short Term Memory," 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), pp. 1234-1237, Feb. 2023, doi: 10.1109/ICIICS59993.2023.10421563.
5. Amalia, A., Sharif, A., Haisar, F., Gunawan, D., & Nasution, B. B. (2016). Meme Opinion Categorization by Using Optical Character Recognition (OCR) and Naïve Bayes Algorithm
6. Y. Ashrita, A. Srinivas, S. Abhiram, P. Rao Vemula, and V. Hemanth, "Deep Learning Techniques for Sentiment Analysis on Social Media Text," 2023 6th International Conference on Contemporary Computing and Informatics (IC3I), pp. 2294-2297, Dec. 2023, doi: 10.1109/IC3I59117.2023.10398065.
7. S. J. Sangatha and Dr. V.M. Aria Anu, "Current Trends on Sentimental Analysis on Youtube Videos," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), pp. 1-4, Dec. 2022, doi: 10.1109/ICOEI53556.2022.9777200.
8. K. Myilvahanan, S. B., T. Raj, C. Attanti, and S. Sahay, "A Study on Deep Learning based Classification and Identification of offensive memes," 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), pp. 1234-1237, Aug. 2023, doi: 10.1109/ICAIS56108.2023.10073933.
9. S. Suryawanshi, B.R. Chakravarthi, M. Arcan, and P. Buitelaar, "Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text," in Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (TRAC-20), Marseille, France, pp. 32-41, May 2020.
10. D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in Advances in Neural Information Processing Systems 34, pp. 2611-2624, 2020. doi:10.48550/arXiv.2005.04790

11. A. Nayak and A. Agrawal, "Detection of hate speech in Social media memes: A comparative Analysis," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), pp. 1234-1237, Dec. 2022, doi: 10.1109/ICICICT54557.2022.9917633.
12. B. Manikandan and Dr. T.K. Kumar, "Analysis of Sentimental on Twitter using Content Aware Support Vector Machines," 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), pp. 1234-1237, Feb. 2023, doi: 10.1109/ICSES60034.2023.10465302.
13. R.K. Giri, S.C. Gupta, and U.K. Gupta, "An approach to detect offence in Memes using Natural Language Processing(NLP) and Deep learning," 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6, Jan. 2021, doi: 10.1109/ICCCI50826.2021.9402406.
14. https://github.com/ilya16/deephumor