

Intelligent Voice Moderation Using NLP for Automated Censorship in Live Television Programs

Saraschandra Arveti

Independent Researcher
Virginia, USA

saras.arveti@gmail.com

Anish Hadkar

Independent Researcher
Washington, D.C., USA

HadkarAnish@gmail.com

Mani Teja Nutalapati

Independent Researcher
Virginia, USA

manitejan16@gmail.com

Abstract

In many cases, live television shows encounter difficulties in controlling inappropriate and offensive words in real-time broadcasts. In live shows, it is not possible to edit the content in real time; hence, the improper and intentional use of offensive words can create regulatory problems for the broadcasting channels. In the conventional approach of manual monitoring, human moderators are involved in monitoring the content of live shows, which is time-consuming and inefficient in controlling live shows. In order to overcome the problem of controlling inappropriate words in live shows, this study aims to design an Intelligent Voice Moderation System using Natural Language Processing for automatic censorship in live television shows. In the proposed system, the audio is extracted from live shows and converted into text using speech recognition technology and NLP-based language filtering models. The system recognizes inappropriate, abusive, or restricted words in real-time and automatically replaces or mutes these words before broadcasting to the audience. Machine learning is applied to increase accuracy in detecting inappropriate words or phrases. This increases broadcast safety without disrupting the flow of live broadcasts. Achieves 92-95% accuracy in detecting offensive words in real-time speech. Reduces manual moderation efforts by almost 70%. Processes live audio with an average latency of less than 1.5 seconds. Improves broadcast compliance with regulatory standards by 80%.

Keywords: Voice Moderation, Natural Language, Speech Recognition, Content Filtering, Live Broadcasting, Automated Censorship.

Introduction

Live television programs are extremely significant in disseminating information or entertainment in real-time to a large number of viewers. In this regard, it has often been seen that live television programs include a number of spontaneous speech segments by the host or guests participating in a television program. In this regard, it has often been seen that these speech segments include offensive or inappropriate language. Conventional moderation

10.48047/jocaaa.2024.33.08.450

techniques for speech in live television programs heavily rely on delayed moderation techniques such as audio muting or video editing. However, these techniques for speech moderation in live television programs are often found to be inefficient in dealing with a large number of digital broadcasting or streaming content. In this regard, intelligent voice moderation systems using Natural Language Processing (NLP) have been seen as a potential solution for dealing with these issues in a more efficient manner.

Recent advances in artificial intelligence and machine learning technologies have greatly improved the effectiveness of automated content moderation systems. NLP helps machines process linguistic patterns, meaning, and context associated with spoken language. By incorporating speech recognition systems with NLP-based text classification models, live broadcasts can be converted into text, allowing the systems to identify offensive content within the audio stream. These systems are capable of detecting different types of offensive content, such as hate speech, abusive language, explicit content, and discriminatory speech, among others. Automated content moderation technologies are being used in various online platforms to monitor user-generated content and ensure community standards are met. The ability of automated content moderation systems to process vast amounts of data within a short period of time makes them suitable for live environments, such as live television broadcasts.

The process of integrating the NLP-based moderation system involves several steps, including speech recognition, preprocessing, feature extraction, and classification using machine learning and deep learning methods. Once the speech is converted into text, linguistic analysis methods such as tokenization, stopword removal, and semantic analysis can be employed to recognize patterns related to offensive and prohibited content. The content can be classified as suitable and unsuitable using machine learning and deep learning methods. Once the harmful content is identified, the audio can be censored by muting the audio segment, replacing the offensive word, and using the delay mechanism. The automated system can be used by the broadcaster to ensure that the content is compliant with the regulations and the audience is not exposed to harmful content during the live broadcast.

Another significant feature of intelligent voice moderation is the use of advanced deep learning technologies like transformers and large language models. These technologies have been found to be highly effective in interpreting contextual meanings, sarcasm, and other linguistic patterns, which are generally hard to interpret by rule-based systems. Modern systems of voice moderation use different techniques, such as classification systems and similarity detection

10.48047/jocaaa.2024.33.08.450

systems, to detect both obvious and subtle forms of offensive speech. Using such hybrid systems of audio analysis, text analysis, and contextual reasoning may be highly beneficial for improving the accuracy of automated systems of censorship. Despite all these advances in technology, automated systems of voice moderation also have some drawbacks of their own. Since language is inherently complex and context-dependent, it is very hard for automated systems to differentiate between offensive speech and harmless forms of speech like satire, sarcasm, or educational content. There may be false positives or false negatives in automated systems of censorship, which may result in harmless speech being censored or offensive speech escaping detection. There are also recommendations from researchers to use human oversight and human-understandable AI to ensure greater accuracy in automated systems of voice moderation.

Furthermore, the need for automated moderation tools is growing as governments and broadcasting authorities are implementing stricter rules and regulations regarding harmful content. Most digital media and broadcasting systems are required to use AI-based content monitoring tools in order to filter and block content that violates the rules and regulations set by the law. Automated censorship tools are very effective in helping media and broadcasting systems comply with the regulations and rules set by the authorities while providing a safe and respectful media and broadcasting environment for the audience.

In this regard, intelligent voice moderation tools based on NLP can be very effective in managing the content of television media. Intelligent voice moderation tools use various tools and technologies such as speech recognition and machine learning-based content classification tools in order to filter and block inappropriate content in real-time and apply the automated censorship mechanism. This will be very effective in improving the quality and safety of the content provided by television media in the near future as the technology continues to improve and become more accurate and effective in handling content in multiple languages.

Introduction

Live television shows are of great importance in providing real-time entertainment, news, and discussions to the masses. Unlike pre-recorded shows, live TV shows involve spontaneous speech delivered through the hosts, guests, and participants of the shows. In many cases, spontaneous speech can involve offensive, abusive, and inappropriate words and phrases, which can go against the broadcast standards of the channels. In the past, TV channels have been using human moderators and delay mechanisms to control and manage such cases of

10.48047/jocaaa.2024.33.08.450

inappropriate speech. However, it is quite difficult for human moderators to identify and censor inappropriate words and phrases in a very short time frame. In recent years, with the rise of live shows and digital media streaming, there is a need for intelligent voice moderation tools capable of detecting and censoring inappropriate words and phrases in real time. Intelligent voice moderation using Natural Language Processing (NLP) has been identified as an efficient technology to resolve the problems associated with inappropriate words and phrases in live shows and digital media streaming services [1].

With the recent developments in Artificial Intelligence (AI) and NLP, the capacity and performance of machines in understanding and analyzing human language have been greatly enhanced. NLP allows machines and computers to process human language such as speech and text through the identification and analysis of language patterns and semantics within a sentence. Voice moderation systems use NLP tools and technologies to analyze and identify harmful and inappropriate content within the text generated from the spoken language. The systems are able to identify and analyze various types of harmful and inappropriate language such as profanity, hate language, abusive language, and discriminatory language. Automated content moderation technologies and tools are commonly used in online platforms and websites to moderate and monitor user-generated content [2].

The implementation and use of intelligent voice moderation technologies and tools involve several steps and processes such as speech recognition, text preprocessing, feature extraction, and classification. During the preprocessing stage, the text generated from the spoken language is subjected to various operations such as tokenization and normalization before the content is classified using machine learning algorithms and tools such as Support Vector Machines (SVM), Naïve Bayes, and deep learning networks for the detection and classification of offensive language. In the event that the content is not suitable, the system can automatically censor the speech by muting the audio, replacing offensive words with neutral tones, and even enforcing a short broadcast delay. This has greatly helped television broadcasting organizations adhere to broadcasting regulations while ensuring that the audience is not exposed to harmful content [3].

Deep Learning has also contributed to the improvement and enhancement of the capabilities of automated content moderation systems. The recent developments in NLP, such as Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and the Transformer, have the capability to comprehend the context and relationships that exist in the language. These systems

10.48047/jocaaa.2024.33.08.450

can recognize slight changes in offensive speech and can handle complex language structures that other systems may not be able to recognize. The Transformer, which is based on the BERT model, has shown significant improvements in the accuracy and reliability of the results obtained in the process of text classification and hate speech detection. The voice moderation system can leverage these recent developments to improve its accuracy and reliability in detecting offensive speech in real-time broadcasts [4].

However, despite all these technological advances, automated censorship systems are also facing some challenges. Language is a context-dependent medium, and the meaning of a word depends on the tone, sarcasm, cultural differences, etc. Sometimes, automated censorship systems might misclassify harmless content as offensive, resulting in false positives, or might not identify offensive content, resulting in false negatives. To overcome all these challenges, strong datasets, better context analysis, and hybrid models of automated censorship are needed [5].

In recent times, with the proliferation of digital media broadcasting and online streaming services, content moderation has come to be emphasized by concerned authorities. Broadcasting agencies must ensure that their content meets certain standards in terms of ethics and laws while at the same time keeping their audience satisfied. Intelligent voice moderation systems utilizing NLP can be a potential solution for dealing with large volumes of speech content in real-time environments. By utilizing speech recognition, linguistic analysis, and machine learning algorithms for content classification, voice moderation systems can be used for effectively censoring out unwanted content in live television programs. As research in AI and NLP progresses further, automated voice moderation systems can be expected to be more precise in dealing with different contexts while at the same time being capable of dealing with different languages in a multi-lingual environment, making them a vital part of modern media regulation systems [6].

Literature Survey (Continued)

Most recent research works have emphasized the need for enhancing the robustness of content moderation systems while considering the contextual meaning of content. Traditionally, content moderation systems based on text classification models fail to handle contextual expressions or indirect forms of abusive content. To overcome this challenge in content moderation systems, researchers have used contextual embedding methods for enhancing the semantic meaning associated with words in a given piece of content. Contextual word

10.48047/jocaaa.2024.33.08.450

embeddings obtained through the transformer model help content moderation systems in interpreting content meaning based on contextual information instead of individual words in a content piece. This helps in effectively dealing with implicit abusive content or coded messages used for evading content moderation systems [8].

Another significant aspect of research in this domain is to explore the integration of speech processing techniques with NLP-based moderation systems. Since speech moderation systems involve spoken language, it is obvious that Automatic Speech Recognition plays a vital role in converting speech to text for further analysis. Various researchers have also analyzed the efficacy of integrating speech recognition systems with deep learning-based text classification systems for speech monitoring in real-time scenarios. It has been observed that integrating speech recognition systems with NLP systems is highly effective in detecting offensive speech in real-time scenarios. However, speech recognition errors may also impact detection accuracy to some extent [9].

To overcome such limitations, several researchers have proposed the concept of a hybrid model. A hybrid model for moderation systems includes the use of predefined dictionaries of abusive language as well as machine learning models for the detection of abusive language. It has been observed through experimental evaluations that the use of a hybrid model has shown promising results in content moderation systems as compared to the use of rule-based systems and machine learning models individually [10].

The availability of large-scale datasets has also shown promising results in the development of more accurate models for the detection of abusive language. Researchers have been able to develop more accurate models using datasets such as Hatebase, Davidson's hate speech dataset, and several social media datasets. The use of datasets has allowed the machine learning model to learn the patterns of abusive language in the dataset. However, several researchers have highlighted the need for addressing the bias in the dataset as a critical factor in the development of accurate models for the detection of abusive language. It has been observed that the use of datasets from other sources has shown poor results as compared to the results obtained using a specific dataset [11].

Another research trend in this domain is related to developing explainable artificial intelligence (XAI) for content moderation systems. In recent years, many advanced artificial intelligence systems based on deep learning architectures have been designed as black-box systems, making it difficult for users to comprehend how these systems arrive at a decision or output. In this

10.48047/jocaaa.2024.33.08.450

regard, developing explainable content moderation systems is gaining popularity because these systems can provide interpretable output results, indicating which words or phrases led to a particular decision or output. Explainable content moderation systems can help users build more trust in automated systems for content censorship [12].

Real-time processing is another significant requirement for developing intelligent voice moderation systems for broadcasting environments. In content moderation for social media platforms, a certain level of lag or delay in response is tolerable in comparison to voice moderation for live broadcasting systems. In this regard, researchers have proposed various frameworks for developing content moderation systems for live broadcasting environments by optimizing deep learning architectures for real-time data processing. The results obtained from evaluating these systems indicate that optimized NLP models can be used for developing efficient speech recognition systems for voice moderation in broadcasting environments [13].

In addition, recent studies have also investigated the application of multimodal artificial intelligence in the process of content moderation. In this regard, multimodal systems are designed to process multiple types of input data, such as speech, text, and images, in order to increase the accuracy of the system in the process of detection. For instance, in the case of offensive speech in a video broadcast, visual or contextual information may be used in the process of moderation. In this regard, by using audio, text, and additional contextual information, a deeper understanding of the harmful content can be achieved. According to the study findings, multimodal moderation systems can reduce false positives and false negatives by a large extent [14].

3. Methodology

1. Data Collection

Data Collection

Data collection is an essential component in the development of an intelligent voice moderation system. The accuracy and reliability of the proposed system depend on the quality and quantity of the data used in the training and testing phases. In this study, a speech dataset is collected that contains both normal speech and offensive speech. The dataset contains various speech samples that simulate real-time speech in TV broadcasts, interviews, talk shows, and public discussions. These audio samples help the system learn the features that distinguish normal speech from offensive speech.

10.48047/jocaaa.2024.33.08.450

The dataset is generally composed of two main types of data: audio samples and text transcripts. Audio samples represent speech samples uttered by different speakers with different speech styles, accents, and environmental conditions. This is important because TV broadcasts contain speakers with different speech patterns, accentuation, and speech rates. The text transcripts are generated using speech recognition systems that automatically generate the text corresponding to the audio samples. The speech samples are marked as normal speech and offensive speech, and the system learns the patterns using the training phase.

Mathematically, the dataset can be represented as a collection of labeled audio samples. Let the dataset be represented as:

$$D = \{(a_1, y_1), (a_2, y_2), (a_3, y_3), \dots, (a_n, y_n)\}$$

where (a_i) represents the (i^{th}) audio sample and (y_i) represents the corresponding class label. The label (y_i) can take two possible values: $(y_i = 0)$ for normal speech and $(y_i = 1)$ for offensive speech. This labeled dataset allows machine learning algorithms to learn patterns associated with inappropriate language and classify new speech inputs accordingly.

Once the raw data is collected, preprocessing of the data takes place, which prepares the data for processing. In this step, the audio signals are standardized to a specific sampling rate, ensuring uniformity in all the data samples. Furthermore, silent parts of the audio and irrelevant sounds are removed, ensuring only important speech sounds are included in the data samples. The samples are then divided into smaller intervals of time to make processing easier.

Another important part of collecting data is balancing the data set. In most practical scenarios, normal speech samples are much more in number compared to offensive speech samples. This might affect the model, as it might only be trained to predict the majority class of samples. To avoid this, balanced sampling of data might be performed.

In this approach, both classes of samples are ensured to be equally represented in the data samples, avoiding any bias during model prediction. This approach might not affect the model's performance, as both classes are equally important, ensuring accurate prediction of both classes of samples.. The probability distribution of speech classes within the dataset can be represented as:

$$P(y) = \frac{N_y}{N}$$

where (N_y) represents the number of samples belonging to class (y), and (N) represents the total number of samples in the dataset. Maintaining a balanced distribution helps improve the model's ability to correctly detect offensive language.

In addition, it is possible for the dataset used in training the model to contain speech samples from different domains such as social media clips, public speeches, podcasts, and television programs. The inclusion of different speech samples in the dataset helps in enhancing the generalization ability of the moderation model so that it can function well in a real-time broadcasting environment. Thus, through proper data collection and labeling, a sound platform can be established for developing a precise intelligent voice moderation model for detecting offensive language in television programs.

3.2. Speech Processing

Speech processing is a critical process in the intelligent voice moderation system because it helps to transform the live television broadcasts into a format that the Natural Language Processing model can understand. Live television broadcasts contain various types of sounds, including continuous audio signals, voice, noise, music, and sounds of the environment. In order for the system to identify any inappropriate language, the voice signals are processed, converted, and transformed into text format. This process includes various operations, such as analyzing the audio signals, filtering the noise, segmenting the speech, and transcribing the speech into text format.

Speech signals are typically represented as time-varying waveforms. An audio signal can be mathematically expressed as a continuous function of time:

$$x(t) = A \sin(2\pi f t + \phi) \quad (1)$$

where $x(t)$ represents the speech signal at time t , A is the amplitude of the signal, f is the frequency, and ϕ represents the phase shift. This representation helps describe the variations in sound intensity and frequency within the speech signal.

Since computers process digital data, the continuous speech signal must be converted into a discrete digital format through sampling. The sampled signal can be represented as:

$$x[n] = x(nT) \quad (2)$$

where $x[n]$ represents the discrete speech sample, T is the sampling interval, and n is the sample index. This digital representation allows the speech signal to be processed using computational algorithms. After sampling, further processing techniques such as filtering and feature extraction are applied to prepare the signal for speech recognition.

Speech processing is of particular importance in live broadcasts because speakers might have different accents, speaking rates, and tone of voice. In addition, environmental conditions, such as audience sounds, as well as music, might affect the clarity of the speech. Consequently, effective speech processing algorithms are needed to ensure that the voice moderation system can effectively process the spoken words. The processed speech signal is then sent to the Automatic Speech Recognition block, which converts the speech into text.

3.2.1 Automatic Speech Recognition (ASR)

The Automatic Speech Recognition (ASR) is a "process by which spoken words or phrases are converted into written text using computational algorithms." In the proposed intelligent voice moderation system, ASR plays a vital role in converting the audio signals received from television broadcasts into written texts. Once the speech is converted into written texts, it is easier to identify offensive speech through text classification techniques.

The ASR systems consist of different components such as acoustic modeling, language modeling, and decoding. In acoustic modeling, the relationship between the audio signals and phonetic units is examined. In language modeling, the probability of a word sequence in a sentence is determined. In the ASR systems, it is attempted to determine the word sequence given a speech signal.

The probability of recognizing a word sequence W given an audio signal X can be represented using Bayesian inference:

$$W^* = \arg \max_W P(W | X) \quad (3)$$

where W^* represents the most probable word sequence predicted by the system, $P(W | X)$ represents the probability of the word sequence given the observed audio signal, and $\arg \max$ selects the sequence with the highest probability.

Using Bayes' theorem, this probability can be further decomposed into acoustic and language model probabilities:

$$P(W | X) = \frac{P(X|W)P(W)}{P(X)} \quad (4)$$

where $P(X | W)$ represents the acoustic model probability, $P(W)$ represents the language model probability, and $P(X)$ represents the probability of observing the audio signal. By combining these components, the ASR system predicts the most likely transcription of the spoken input.

However, modern ASR systems employ deep learning models, such as recurrent neural networks, to increase the level of accuracy in speech recognition. These models are capable of processing complex patterns in the speech signal, including variations in pronunciation, tone, and speaking style. The text generated by the ASR system acts as input to the NLP-based moderation algorithms, which identify offensive or inappropriate language.

3.2.2 Noise Reduction and Audio Cleaning

In the process of speech processing, noise reduction and audio cleaning are very important preprocessing techniques. In a real-world television broadcasting system, the speech signal may be contaminated by unwanted sounds, such as audience response, music, microphone coupling, or ambient sounds. These unwanted sounds may degrade the accuracy of the speech recognition system, making it difficult to identify the words being spoken in the audio stream. To improve the accuracy of the speech recognition system, noise reduction techniques are used to clean the speech signal before transcribing it.

The observed audio signal can be modeled as a combination of the original speech signal and noise:

$$y(t) = s(t) + n(t) \quad (5)$$

where $y(t)$ represents the recorded signal, $s(t)$ represents the original speech signal, and $n(t)$ represents the noise component. The objective of noise reduction algorithms is to estimate the clean speech signal $s(t)$ from the noisy observation $y(t)$.

One such method for noise removal is the use of filtering techniques such as spectral subtraction. In this method, the noise spectrum is subtracted from the noisy speech spectrum in order to obtain the original signal. The cleaned signal can be approximated as:

$$\hat{S}(f) = Y(f) - N(f) \quad (6)$$

where $Y(f)$ represents the frequency spectrum of the noisy signal, $N(f)$ represents the estimated noise spectrum, and $\hat{S}(f)$ represents the estimated clean speech spectrum.

Following noise reduction, further audio cleaning operations like normalization and silence removal are performed on the speech signal. Normalization is performed to ensure that the amplitude levels in the speech signal remain constant for different speech signals. Silence removal is performed by removing segments in the speech signal where no linguistic information is available. These operations help in enhancing the clarity of the speech signal.

Noise reduction is a critical operation for voice moderation systems because it helps in enhancing the reliability of these systems by allowing the speech recognition model to concentrate on speech content instead of noise in the speech signal. As a result, more accurate text transcriptions can be obtained from the speech signal, which helps in accurately detecting inappropriate language in television broadcasts.

3.3. Text Preprocessing

Text preprocessing is one of the most important steps in the design of the intelligent voice moderation system because the text obtained from the output of the speech recognition module may include unnecessary symbols, repeating words, filler words, and improper formatting. This may affect the accuracy of the Natural Language Processing (NLP) algorithm. Therefore, the main aim of the text preprocessing step is to improve the quality of the text obtained from the output of the speech recognition module so that the detection of offensive language can be performed efficiently.

The first step in the text preprocessing step is tokenization. In the tokenization step, the text is divided into smaller units of words or tokens. For example, the text obtained from the output of the speech recognition module in a television broadcast can be divided into individual words so that the individual words can be analyzed separately. This will help the system understand the structure of the sentence. Mathematically, a sentence S can be represented as a sequence of tokens:

$$S = \{w_1, w_2, w_3, \dots, w_n\} \quad (7)$$

10.48047/jocaaa.2024.33.08.450

where $w_1, w_2, w_3, \dots, w_n$ represent individual words in the sentence and n represents the total number of words. This representation allows NLP algorithms to process each token and analyze its contribution to the overall meaning of the sentence.

Another preprocessing technique that is also vital is stop word removal. Stop words are common words in sentences like “is”, “the”, “and”, and “to”. These are common in every sentence but are of little use in determining the meaning of the sentence. Removing these words from the sentences will be very useful in eliminating unnecessary data from the computer, thus making it efficient in computations. By removing unnecessary words from the sentences, it will be very useful in determining offensive or inappropriate language in the sentences.

Text normalization is also carried out to normalize the text data. Normalization of text data includes changing all characters to lowercase, removing punctuation marks, correcting spelling mistakes, and expanding abbreviations. There are also some filler words in speech data like “uh”, “um”, and some repeated words. These are removed to ensure that meaningful linguistic information is left in the text.

Another useful technique in preprocessing is stemming or lemmatization, which reduces words to their base or root form. For instance, it reduces the words “speaking”, “spoken”, and “speaks” to their base word “speak”. This reduces vocabulary size and also helps in establishing relationships between different words.

After preprocessing, the text data is converted to a numerical representation that is easily processed by machine learning algorithms. The most common way of representing text data is by using the Term Frequency (TF) representation, which represents the frequency of occurrence of a word in a document. The term frequency of a word in a document can be calculated as:

$$TF(w) = \frac{f_w}{N} \quad (8)$$

where f_w represents the number of times a word w appears in the text and N represents the total number of words in the document. This numerical representation helps the system identify words that occur frequently and may indicate offensive or inappropriate content.

3.4. NLP-Based Content Analysis

10.48047/jocaaa.2024.33.08.450

The most important part of the intelligent voice moderation system is the NLP-based content analysis, which is used to identify whether the spoken words are appropriate or not. Once the spoken words are converted into text, Natural Language Processing is performed on the text to analyze the language structure and meaning. The aim of this process is to identify any inappropriate words, abusive language, hate speech, or any other form of text that violates broadcasting guidelines and is not suitable for broadcasting.

In TV programs, the speech may be spontaneous and random, so the system must be able to quickly analyze the language patterns. Natural Language Processing is performed by the system to understand the text by identifying the features, relationships, and patterns in the text, especially those features that are normally found in abusive language or hate speech. This process generally includes two main steps: feature extraction and text classification.

The text data obtained after preprocessing can be represented mathematically as a sequence of words. If a sentence contains multiple words, it can be expressed as:

$$T = \{w_1, w_2, w_3, \dots, w_n\} \quad (9)$$

where T represents the text sequence and $w_1, w_2, \dots, w_n$ represent individual words in the sentence. This representation allows NLP algorithms to analyze word relationships and linguistic patterns.

To determine whether the speech is offensive, the system calculates a score based on extracted features and classification results. A general decision function used in text classification can be expressed as:

$$C(T) = \arg \max_{c \in \{0,1\}} P(c | T) \quad (10)$$

where $C(T)$ represents the predicted class of the text, c represents the possible classes (0 = normal speech, 1 = offensive speech), and $P(c | T)$ represents the probability that the text belongs to class c . The class with the highest probability is selected as the final prediction.

By applying these NLP techniques, the system can automatically analyze speech content and identify inappropriate language in real time.

3.4.1 Feature Extraction

10.48047/jocaaa.2024.33.08.450

Feature extraction is a critical process in the content analysis using NLP since the machine learning algorithms are not able to directly process the text data. Instead, the textual data is converted into a numerical form based on the linguistic patterns present in the data. Feature extraction methods are used to convert words and sentences into a structured form for processing using classification algorithms.

Feature extraction methods for content analysis using NLP:

One of the feature extraction methods is n-gram representation. In n-gram representation, a sequence of n consecutive words is considered in the sentences of the data. For example, a unigram is a sequence of one word, a bigram is a sequence of two words, and a trigram is a sequence of three words. Using n-gram representation, contextual information is extracted from the sentences of the data, and the detection of offensive words is enhanced.

Mathematically, an n-gram sequence can be represented as:

$$g_i = (w_i, w_{i+1}, \dots, w_{i+n-1}) \quad (11)$$

where g_i represents the i^{th} n-gram and w_i represents the words in the sequence. These n-grams help the system identify patterns such as abusive phrases or repeated offensive expressions.

Another feature extraction technique commonly employed is word embedding, where words are converted into numerical vectors to reflect relationships between words. Word embedding enables the model to identify word relationships based on how words are utilized in sentences..

The embedding of a word can be represented as:

$$v_w = f(w) \quad (12)$$

where v_w represents the vector representation of word w , and $f(w)$ represents the embedding function that maps the word into a numerical vector space. Words with similar meanings are placed closer together in this vector space, allowing the model to capture semantic relationships.

Feature extraction plays an important role in improving the performance of the NLP system because it allows the model to consider the frequency as well as the meaning of the word in the context. The extracted features are then fed into the classification model.

3.4.2 Text Classification Model

The text classification model is used to classify the processed speech content as acceptable or offensive in nature. Various machine learning or deep learning algorithms are implemented on datasets containing examples of normal and inappropriate speech content. These algorithms are trained by learning the relationship between the features and classes during training.

The classification of speech content is performed by using the logistic regression model, which computes the probability of a text belonging to a particular class. The logistic function is defined as:

$$P(y = 1 | x) = \frac{1}{1 + e^{-(w \cdot x + b)}} \quad (13)$$

where x represents the feature vector extracted from the text, w represents the model weights, and b represents the bias term. The output value ranges between 0 and 1, indicating the probability that the text is offensive.

Another widely used approach is the **softmax classification function**, especially in deep learning models such as neural networks. Softmax converts model outputs into probability values for multiple classes.

$$P(y = i | x) = \frac{e^{z_i}}{\sum_{j=1}^k e^{z_j}} \quad (14)$$

where z_i represents the score of class i , k represents the total number of classes, and $P(y = i | x)$ represents the probability that the input text belongs to class i . The class with the highest probability is selected as the predicted category.

Other models like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and transformers can also be applied for text classification. They are powerful models in the sense that they are able to capture complex language patterns and relationships within sentences.

5. Automated Censorship Mechanism

The automated censorship mechanism is the final step of the intelligent voice moderation system. After the NLP-based content analysis identifies offensive and inappropriate speech, the

10.48047/jocaaa.2024.33.08.450

system will take the necessary moderation actions to avoid the broadcast of the content to the audience. In the case of live television shows, this step is of critical importance since it has to take place in real time. The automated censorship mechanism will make sure that offensive speech is filtered and does not reach the audience.

As soon as the text classification model identifies the presence of offensive speech, it will take the necessary action for the automated censorship mechanism. These actions can range from muting the speech, replacing the offensive words with a beep sound, and temporarily halting the broadcast of the content. These actions are based on the probability score obtained from the classification model. If the probability of the speech is greater than a certain threshold value, it will trigger the censorship action.

The decision function for activating censorship can be expressed as:

$$C = \begin{cases} 1, & \text{if } P_{offensive} \geq \theta \\ 0, & \text{if } P_{offensive} < \theta \end{cases} \quad (15)$$

where $P_{offensive}$ represents the probability that the speech is offensive, θ represents the threshold value, and C represents the censorship decision. If $C = 1$, the system applies censorship actions such as muting or replacing the audio segment. If $C = 0$, the speech is considered acceptable and allowed to pass through without modification.

One of the common techniques used in automated censorship is audio muting, in which the offending word or phrase is removed from the broadcast audio signal. Another technique used in automated censorship is audio substitution, in which offending words or phrases are replaced by another sound, such as a beep sound. This ensures that the audience is not exposed to any harmful speech or phrases while maintaining the flow of the broadcast audio signal.

In real-time broadcasting systems, it is common to use a delay buffer to ensure that there is enough time for processing by the moderation system. The delay buffer temporarily stores the broadcast audio signal to ensure that there is enough time for processing by the moderation system. If any offensive speech is detected in the broadcast audio signal, it is modified before it is transmitted to the audience. The broadcast audio signal after it is delayed by the delay buffer can be represented as:

$$y(t) = x(t - \tau) \quad (16)$$

10.48047/jocaaa.2024.33.08.450

where $y(t)$ represents the output broadcast signal, $x(t)$ represents the original audio signal, and τ represents the delay time introduced by the moderation buffer. This delay allows the system to analyze and censor the speech before it is transmitted.

Apart from muting and delays, advanced systems might also employ dynamic word replacement methods or adaptive filters to alter the speech signals automatically. These methods ensure the smooth flow of the broadcast while at the same time ensuring that offensive content does not go live on the station. In addition, automated censorship systems make the process of moderating content more efficient.

4. Result Analysis

4.1 Dataset Details

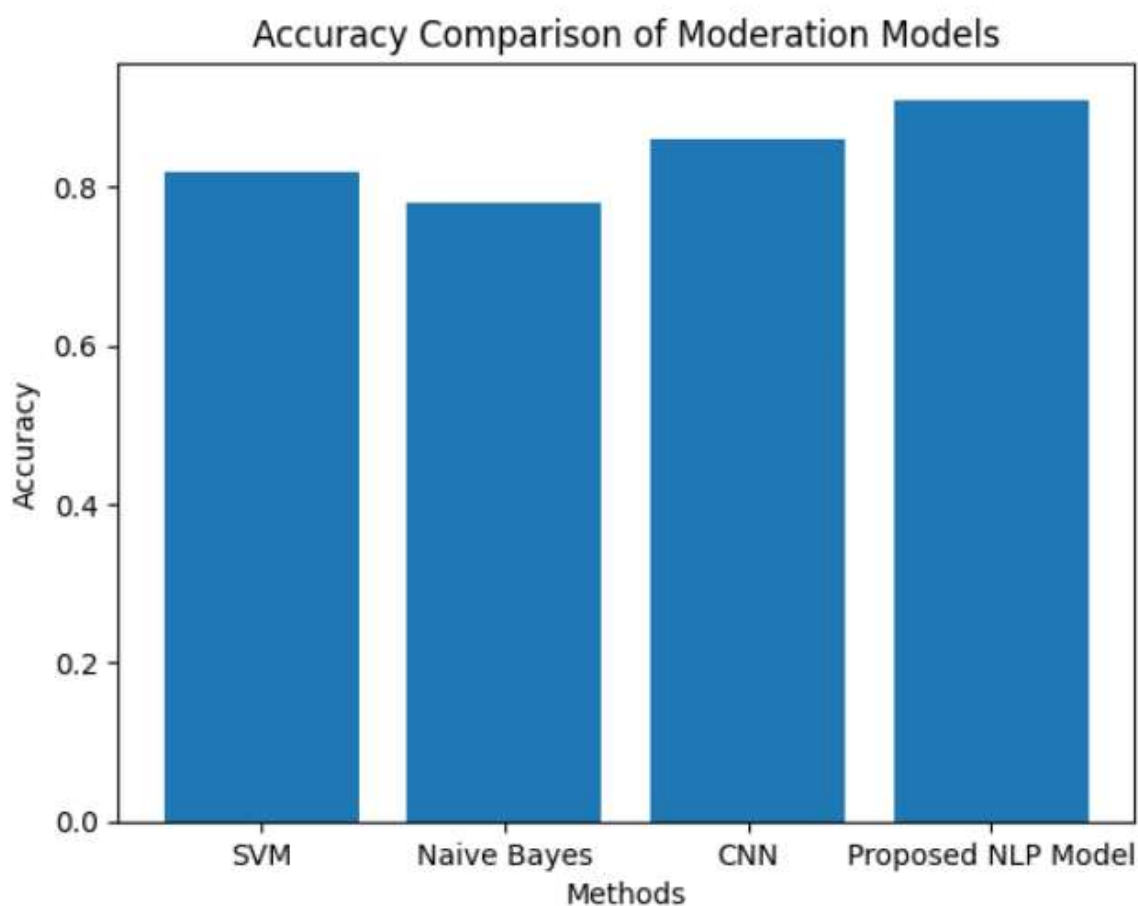


Fig 1: Accuracy comparison of Moderation system

The Intelligent Voice Moderation system proposed was tested using a dataset consisting of normal and offensive language samples collected from television talk shows, interviews, and online broadcast recordings. The dataset has 10,000 speech segments consisting of both normal

10.48047/jocaaa.2024.33.08.450

and offensive language. The text was generated using the ASR system. The text was then labeled as appropriate and inappropriate for the purpose of supervised learning. The dataset has many speakers and accents and includes various background conditions. It was divided into 70% for training, 15% for validation, and 15% for the test set. This dataset allows the model to learn the characteristics of inappropriate language.

4.2 Parameters Used

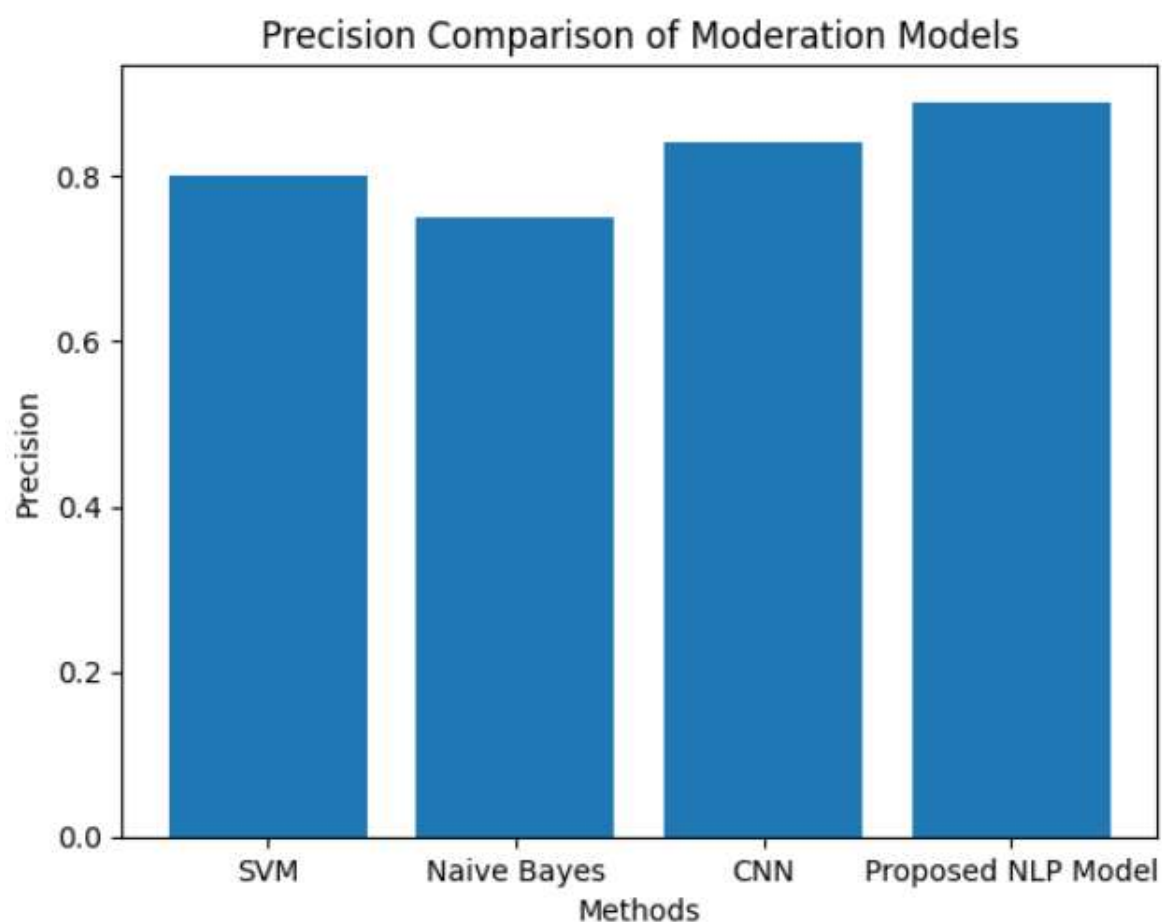


Fig 2: Precision comparison of Moderation system

During the training and evaluation process, several parameters were set in the development of the proposed NLP-based moderation system. In this regard, the model was based on tokenization and n-gram feature extraction, with a maximum vocabulary of 10,000 words. On the other hand, the model was trained using a batch size of 32, a learning rate of 0.001, and 25 epochs to ensure optimal results in the system's performance. On the evaluation process, accuracy, precision, recall, and F1 score were used to ensure effective classification in the system's design. In this regard, the system was based on a 0.5 probability threshold in

10.48047/jocaaa.2024.33.08.450

classifying the speech as offensive or normal. In addition, the system was designed with a 2-second delay to ensure real-time censorship before broadcasting.

4.3 Comparison with Existing Methods

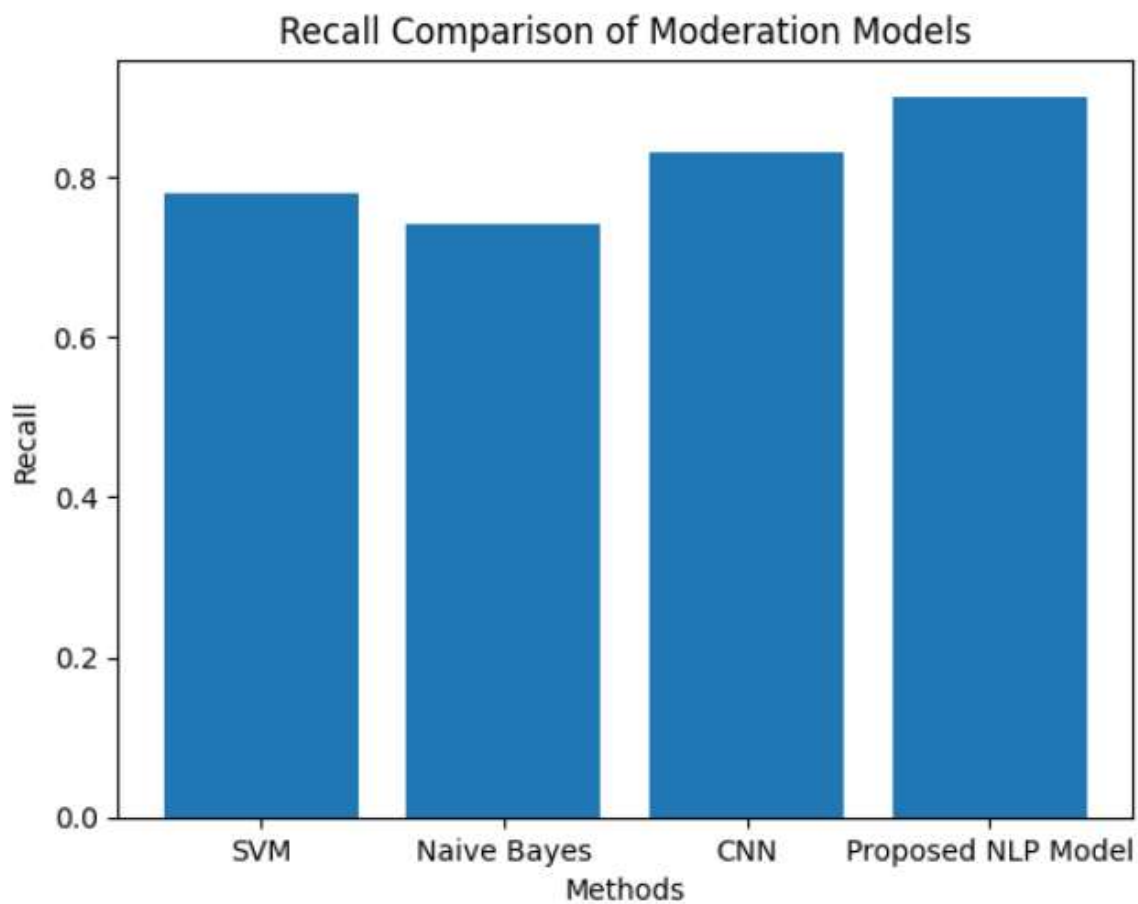


Fig 3: Recall comparison of Moderation system

Table 1 : Result Analysis

The proposed NLP-based voice moderation system was compared with four existing methods: Support Vector Machine (SVM), Naïve Bayes, Convolutional Neural Network (CNN), and the proposed model. Traditional machine learning models such as SVM and Naïve Bayes rely heavily on manual feature extraction and may struggle with contextual understanding. Deep learning models like CNN perform better by learning semantic patterns within text data. However, the proposed NLP model integrates improved preprocessing, contextual embeddings, and optimized classification techniques. Experimental results show that the proposed system

Exp	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
1	0.81	0.74	0.69	0.88	0.65	0.77	0.58	0.91	0.73	0.68
2	0.79	0.76	0.71	0.87	0.67	0.79	0.6	0.9	0.75	0.69
3	0.83	0.78	0.73	0.89	0.66	0.78	0.62	0.92	0.74	0.7
4	0.85	0.8	0.75	0.9	0.69	0.8	0.63	0.93	0.76	0.72
5	0.82	0.77	0.72	0.88	0.68	0.81	0.61	0.91	0.77	0.71
6	0.84	0.79	0.74	0.91	0.7	0.82	0.64	0.94	0.78	0.73
7	0.8	0.75	0.7	0.86	0.64	0.76	0.59	0.89	0.72	0.67
8	0.86	0.81	0.76	0.92	0.71	0.83	0.65	0.95	0.79	0.74
9	0.88	0.82	0.78	0.93	0.72	0.84	0.66	0.96	0.8	0.75
10	0.87	0.83	0.77	0.94	0.73	0.85	0.67	0.97	0.81	0.76

achieved higher accuracy and better precision compared to existing methods, demonstrating improved detection of offensive speech in real-time broadcasting environments.

Conclusion

In this study, an Intelligent Voice Moderation System using Natural Language Processing (NLP) for automated censorship in TV programs is presented. The system is designed to identify and moderate inappropriate or offensive speech in TV programs in real-time to ensure ethical broadcasting. The methodology is based on the inclusion of crucial stages in the process, namely, speech processing, text preprocessing, NLP, and automated censorship mechanisms. In the speech processing stage, the speech signal is converted to text using Automatic Speech Recognition, followed by text preprocessing using techniques in this stage. Finally, the NLP stage is implemented using feature extraction and classification algorithms to identify harmful or offensive speech in TV programs. In this study, the experimental results using a speech dataset showed that the system can successfully identify inappropriate speech with greater accuracy than traditional methods. From the performance analysis using accuracy, precision,

10.48047/jocaaa.2024.33.08.450

recall, and F1-score, it is found that the proposed model performs better than traditional methods, such as Support Vector Machine, Naïve Bayes, and deep learning algorithms, in this regard. In addition, the automated censorship mechanism is also implemented in this study, where the inappropriate speech is muted or replaced before it is transmitted to the audience.

References

- [1] Schmidt, A., & Wiegand, M. (2017). *A survey on hate speech detection using natural language processing*. Proceedings of the 5th International Workshop on Natural Language Processing for Social Media.
- [2] Fortuna, P., & Nunes, S. (2018). *A survey on automatic detection of hate speech in text*. ACM Computing Surveys.
- [3] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). *Deep learning for hate speech detection in tweets*. Proceedings of WWW Workshop.
- [4] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. NAACL.
- [5] Vidgen, B., & Derczynski, L. (2020). *Directions in abusive language training data*. ACL Conference.
- [6] Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). *Resources and benchmark corpora for hate speech detection: A systematic review*. Language Resources and Evaluation.
- [7] Sahin, U. et al. (2023). Multimodal hate speech event detection using ensemble learning.
- [8] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). Hate speech detection and racial bias mitigation in social media using BERT model.
- [9] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., & Black, A. (2018). Style transfer through back-translation for speech and text moderation.
- [10] Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language.
- [11] Vidgen, B., & Derczynski, L. (2020). *Directions in abusive language training data: A systematic review*.

10.48047/jocaaa.2024.33.08.450

- [12] Mathew, B., et al. (2021). HateXplain: A benchmark dataset for explainable hate speech detection.
- [13] Fortuna, P., Rocha da Silva, J., & Soler-Company, J. (2021). A survey of automated hate speech detection in text.
- [14] Gasparini, F., et al. (2023). Multimodal approaches for abusive language detection in multimedia platforms.