Om Agarwal, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

CleanTalk: An NLP-Based Framework for Profanity Detection and Severity Analysis in Video Content

Om Agarwal[1], Dr. Jasbir Kaur[2], Ifrah Kampoo[3]
Student of Master of Computer Applications (MCA),Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[1]
Director, GNIMS B-School, Head of Information Technology and HR,Guru Nanak Institute of Management Studies, Matunga,Mumbai,India[2]
Assistant Professor, Guru Nanak Institute of Management Studies, Matunga, Mumbai, India[3]

Abstract- This paper presents CleanTalk, a Natural Language Processing (NLP)-driven system designed to automate the detection and severity assessment of profane language within video content. The system operates by extracting audio tracks from video files, transcribing the resulting audio into textual form using speech-to-text technology, and subsequently applying NLP-based analytical techniques to identify profane words and evaluate their severity. Each detected instance is annotated with a corresponding timestamp, and the system produces an overall severity score summarizing the nature of the language throughout the content. The final output consists of an annotated video file with visual markers at profanity timestamps alongside a structured severity report. A pilot evaluation was conducted across three video samples of varying content type, yielding an average precision of 96.7% and an average recall of 84.5%, with performance degradation observed primarily in multi-speaker and fast-speech scenarios. CleanTalk demonstrates strong potential as a scalable, automated solution for content moderation workflows on digital platforms.

Keywords— Natural Language Processing, speech-to-text, profanity detection, content moderation, severity analysis, video annotation.

I. INTRODUCTION

1. Background

The rapid proliferation of user-generated content across major digital platforms — including YouTube, TikTok, and Instagram — has placed significant strain on content moderation systems. Ensuring that uploaded material adheres to community standards and regulatory guidelines has emerged as a critical operational challenge for these platforms. Among the most prevalent policy violations is the use of offensive and profane language, which appears frequently in videos across diverse genres and demographics. Manual review of such content is both resource-intensive and poorly scalable, necessitating the development of robust automated moderation tools.

2. Research Problem

Existing profanity detection solutions are largely designed to process textual input and generally lack the capacity to analyze audio streams derived from video content. Furthermore, most current tools do not incorporate context-aware severity assessment, treating all instances of offensive language uniformly regardless of frequency, linguistic intensity, or contextual usage. This gap highlights the need for a comprehensive system capable of operating directly on video audio while providing nuanced severity evaluations.

3. Problem Statement

The absence of an integrated, audio-capable profanity detection framework that simultaneously performs context-sensitive severity analysis represents a significant limitation in automated video content moderation. This study addresses that gap by proposing CleanTalk, a system that processes audio extracted from video files, identifies profane language through NLP techniques, and produces

severity-annotated outputs suitable for moderation workflows.

4. Objectives

The primary objectives of this study are as follows:

To extract audio streams from video files and convert them into a format suitable for speech recognition processing.

To accurately transcribe the extracted audio into text using speech-to-text technology.

To identify profane words within the transcribed text using NLP-based techniques and predefined lexicons.

To assign severity ratings to each identified instance based on frequency, context, and linguistic intensity.

To generate annotated video outputs with timestamped markers and a consolidated severity report.

5. Research Question

How can Natural Language Processing and automatic speech recognition be effectively leveraged to identify and assess the severity of profane language present in the audio tracks of video content?

6. Scope

The scope of this study is confined to the analysis of audio content extracted from video files. Visual data present within the video is not subject to analysis. The system is designed and evaluated exclusively for English-language content.

II. LITERATURE REVIEW

1. Existing Approaches in Profanity Detection

Prior research in automated profanity detection has predominantly focused on keyword-based filtering methodologies and machine learning classifiers applied to textual data. While these approaches have demonstrated reasonable performance in text-based moderation tasks, they are fundamentally limited in their applicability to audio and video environments. Additionally, earlier models have often failed to account for the contextual nuances inherent in natural speech, including sarcasm, irony, and culturally specific slang, which can significantly alter the offensiveness of a given utterance [1].

2. NLP in Audio and Video Analysis

The integration of NLP methodologies with speech recognition technologies has gained considerable momentum in recent years. Established libraries such as spaCy [3] and NLTK, combined with cloud-based speech recognition services such as the Google Speech-to-Text API [2], have made it increasingly feasible to extract structured, semantically meaningful text from continuous audio streams. This convergence of NLP and speech recognition lays the technical foundation for systems such as CleanTalk, which require accurate transcription as a prerequisite for linguistic analysis.

3. Challenges in Profanity Detection and Severity Assessment

One of the most persistent challenges in this domain is the context-sensitivity of offensive language. A term that constitutes a profanity in one context may be entirely neutral or even affirmative in another, depending on speaker intent, social setting, and audience. Tone, sarcasm, and regional slang further complicate accurate detection, and existing rule-based systems frequently struggle to handle such variability. Additionally, speech-to-text transcription errors introduce a secondary layer of uncertainty that can adversely affect downstream NLP processing [4].

III. METHODOLOGY

1. Research Design

This study adopts an experimental research design. Audio is systematically extracted from video files, converted into a text representation through automated transcription, and subsequently analyzed for profane content using a combination of predefined lexicons and NLP-based processing techniques. The overall pipeline is evaluated on publicly available video samples containing varying degrees and types of profane language.

2. Materials and Tools

The following materials and software tools constitute the primary resources employed in this study:

- Video Sources: Publicly available video files exhibiting varying degrees of profane language, including political commentary compilations, feature film excerpts, and live-stream recordings.
- Speech-to-Text Engine: Google Speech-to-Text API [2], used for converting extracted audio into textual transcripts.
- NLP Libraries: Python-based libraries, specifically spaCy [3] and NLTK, employed for tokenization, linguistic filtering, and profanity identification.
- Profanity Lexicons: Predefined word databases used as reference corpora for identifying offensive terms within transcripts.
- Audio Processing: FFmpeg, used for extracting and converting audio streams from video files into the WAV format.

3. Experimental Setup

The experimental environment involves processing video files through a sequential pipeline implemented in Python. FFmpeg extracts the audio track from each input video and encodes it as a WAV file. The resulting audio is passed to the Google Speech-to-Text API, which returns a time-aligned textual transcript. The transcript is subsequently processed using spaCy and NLTK for tokenization and term identification, after which detected terms are cross-referenced against the profanity lexicon.

Three video samples were selected to represent distinct real-world content categories: a political speech compilation featuring a single dominant speaker, a dramatic film scene involving multiple simultaneous speakers, and a live-stream recording characterized by rapid and informal speech patterns. Ground truth profanity counts were established through manual review prior to automated processing.

4. Data Processing Pipeline

The end-to-end data processing workflow consists of the following sequential steps:

- Audio Extraction: FFmpeg is used to extract the audio track from each input video and encode it as a WAV file, ensuring compatibility with the subsequent transcription component.
- Speech Transcription: The extracted WAV audio is passed to the Google Speech-to-Text API, which produces a time-aligned textual transcript of the spoken content.
- Text Processing: The transcribed text undergoes tokenization and linguistic filtering using spaCy and NLTK. Tokens are cross-referenced against a profanity lexicon to identify candidate offensive terms.
- Severity Analysis: Each identified profane word is assigned a severity rating based on a rule-based classification scheme that considers frequency of occurrence, contextual positioning, and the inherent linguistic intensity of the term.

5. Severity Scoring System

Profane words detected by the system are classified into one of three severity tiers: mild, moderate, or severe. This categorization is performed by a rule-based classifier that evaluates each term along three dimensions: its absolute frequency of occurrence within the video, its contextual usage as inferred from surrounding tokens, and its linguistic intensity as encoded within the reference lexicon. The individual severity ratings are subsequently aggregated to produce an overall severity score for the video, expressed on a normalized scale.

6. Output Generation

The system produces the following outputs upon completion of the analysis pipeline:

- A timestamped list of all profane words detected within the video, including their individual severity classifications.
- An overall severity score summarizing the cumulative level of offensive language present in the content.
- An annotated version of the original video, incorporating visual alert markers at the precise timestamps where profane language was identified.

IV. RESULTS

1. Detection Performance Across Video Samples

Table I presents the detection outcomes for each of the three video samples evaluated in this study. For each sample, the table reports the number of profane instances detected by CleanTalk, the ground-truth count established through manual

review, the number of instances missed, the severity distribution across the three tiers, and the overall severity score assigned by the system. Video 3 returned 20 detected instances against a manually verified ground truth of 18, indicating 2 false positives attributable to transcription misidentification rather than missed detections. Recall exceeds 100% in Video 3 due to 2 false positive detections; effective recall (true positives / ground truth) is 100% as all 18 actual instances were captured.

Table I: Detection Results and Severity Distribution Across Evaluated Videos

Video	Detected	Actual	Missed	Mild	Moderate	Severe	Score
Video 1 – Trump Compilation	50	55	5	20	20	10	2.5
Video 2 – Wolf of Wall Street	25	40	15	10	10	5	1.7
Video 3 – Livestream Compilation	20	18	−2*	10	7	1	1.2

2. Transcription Quality Observations

Transcription performance varied substantially across the three test videos in a manner directly correlated with audio characteristics. Video 1, featuring a single dominant speaker in a controlled broadcast setting, produced clean, accurate transcriptions with no notable errors, enabling the system to detect 50 of 55 profane instances (a miss rate of approximately 9.1%). The five missed instances were attributable to audio compression artifacts rather than transcription failure.

Video 2, an excerpt from a feature film involving multiple simultaneous speakers with overlapping dialogue and ambient soundtrack audio, yielded considerably more transcription errors. The Google Speech-to-Text API struggled to isolate individual voices, resulting in incomplete transcription segments and a detection count of only 25 out of 40 actual instances — a miss rate of 37.5%. This represents the most significant performance degradation observed across the evaluation.

Video 3, a live-stream recording featuring a single speaker delivering rapid, informal speech, produced a modest number of transcription inaccuracies. Fast delivery caused certain utterances to be truncated or phonetically misrepresented in the transcript. Despite this, CleanTalk detected all 18 actual profane instances (with 2 additional false positives), demonstrating strong lexicon matching even under degraded transcription conditions.

3. Precision, Recall, and F1 Scores

Table II summarizes the precision, recall, and F1 scores computed for each video sample. Precision measures the proportion of system-flagged instances that constitute genuine profanities, while recall measures the proportion of actual profanities that the system successfully detected.

Table II: Precision, Recall, and F1 Score by Video Sample

Video	Precision (%)	Recall (%)	F1 Score
Video 1 – Trump Compilation	100.0	90.9	0.952
Video 2 – Wolf of Wall Street	100.0	62.5	0.769
Video 3 – Livestream Compilation	90.0	100.0**	0.947
Average	96.7	84.5	0.889

4. Severity Score Summary

Video 1 received the highest overall severity score of 2.5, consistent with its dense distribution of profane content and the relatively high proportion of moderate and severe instances (20 and 10 respectively). Video 2 was assigned a score of 1.7, reflecting a moderate content profile with 10 mild and 10 moderate instances among the 25 detected. Video 3 received the lowest severity score of 1.2, corresponding to a predominantly mild distribution (10 mild, 7 moderate, 1 severe) across its 20 detected instances.

V. DISCUSSION

1. Interpretation of Results

The pilot evaluation of CleanTalk demonstrates that the system is capable of detecting profane language in video audio with strong precision across all three test scenarios. The average precision of 96.7% indicates that the lexicon-based matching approach generates very few false positives under normal operating conditions, with the minor exception observed in Video 3 arising from phonetic ambiguity in fast-speech transcription rather than a fundamental flaw in the matching logic.

Recall performance, however, proved significantly more sensitive to audio conditions. The system achieved 90.9% recall in the controlled single-speaker environment of Video 1, but dropped to 62.5% in the multi-speaker film excerpt of Video 2. This gap highlights that the primary bottleneck in CleanTalk's detection pipeline is not the NLP analysis layer but rather the upstream speech-to-text transcription. When the transcription is accurate and complete, the profanity identification component performs reliably; when transcription quality degrades — due to overlapping voices, background noise, or rapid delivery — detection coverage suffers proportionally.

The severity scoring mechanism produced intuitively consistent results across the three videos. The political speech compilation, which contained the highest density and intensity of profane language, received the highest severity score. The live-stream recording, where profanity was sparse and predominantly mild, received the lowest. This monotonic relationship between content characteristics and severity scores suggests that the rule-based classifier is functioning as intended within the bounds of this preliminary evaluation.

2. Comparison with Existing Approaches

Existing text-based profanity filters, which operate directly on user-submitted captions or text inputs, would be expected to achieve higher recall in controlled conditions owing to the absence of transcription error. However, such tools are entirely inapplicable to video content that lacks accompanying transcripts — a scenario that constitutes the majority of user-generated video uploads. CleanTalk addresses this gap by integrating audio extraction and transcription as upstream stages, extending profanity detection capability to the raw video format at the cost of some recall in acoustically challenging conditions [1].

The performance profile observed in this study — high precision with variable recall — is consistent with characteristics reported in prior work on speech-based content analysis, where transcription quality is identified as the dominant factor governing end-to-end system accuracy [4]. CleanTalk's results affirm this finding and underscore the importance of transcription robustness as a priority for future development.

3. Practical Implications for Content Moderation

The results of this evaluation indicate that CleanTalk is best suited, in its current form, for content featuring a single speaker in a relatively clean acoustic environment — a category that encompasses a significant proportion of user-generated video on platforms such as YouTube and TikTok, including commentary videos, vlogs, and spoken editorial content. For such content, the system provides a reliable first-pass moderation signal that can substantially reduce the manual review burden on human moderators.

For acoustically complex content — such as multi-person discussions, reaction videos, or content with heavy background audio — the system's lower recall necessitates supplementary human review. Deploying CleanTalk as a triage layer that flags high-confidence detections while routing uncertain cases to human reviewers would constitute an operationally sound moderation workflow given the current performance profile.

Limitations

Several limitations have been identified that constrain the system's performance in certain operational contexts:

- Speech-to-Text Accuracy: The quality of profanity detection is directly dependent on the accuracy of upstream transcription. As demonstrated in Video 2 and Video 3, overlapping speech and rapid delivery can cause significant transcription degradation, resulting in missed detections. This limitation is partly inherent to current speech recognition technology rather than to CleanTalk's own logic.
- Absence of Visual Context: The system analyzes only the audio stream of a video and does not incorporate any visual information, which limits its ability to assess profanity conveyed through gestures, text overlays, or symbolic imagery.
- Restricted Language Support: In its current form, CleanTalk is designed exclusively for English-language content and does not support multilingual or code-switching speech.
- Rule-Based Severity Classification: The current severity scoring system relies on a predefined rule-based classifier rather than a learned model, which may limit its ability to capture nuanced contextual differences in how the same word is used across different video types.
- Small Evaluation Sample: The pilot evaluation was conducted across only three video samples. A more extensive evaluation across a larger and more diverse dataset would be necessary to draw statistically robust conclusions about system performance.

VI. CONCLUSION

This paper has introduced CleanTalk, an NLP-driven framework designed to automate the detection and severity assessment of profane language within video content. The system integrates audio extraction via FFmpeg, automatic transcription through the Google Speech-to-Text API, and NLP-based analysis using spaCy and NLTK to produce timestamped profanity reports, severity scores, and annotated video outputs.

A pilot evaluation conducted across three videos of distinct content types yielded an average precision of 96.7% and an average recall of 84.5%, with an overall F1 score of 0.889. Results indicate that the system performs most reliably in single-speaker, acoustically clean environments and experiences recall degradation in multi-speaker and fast-speech scenarios due to upstream transcription limitations. Despite these constraints, CleanTalk demonstrates meaningful potential as a scalable first-pass moderation tool capable of reducing the manual review burden on content platform moderators.

Future refinements, particularly the integration of more robust transcription models and context-aware deep learning classifiers, are expected to further enhance system performance across a wider range of video types.

Future Work

- Several directions for future development have been identified to extend the capabilities of CleanTalk:
- Integration of deep learning language models, such as BERT, to enable context-aware profanity detection that moves beyond rule-based lexicon

matching and can handle sarcasm, slang, and evolving offensive terminology.

- Adoption of more robust, speaker-diarization-capable speech recognition models to improve transcription accuracy in multi-speaker and overlapping-dialogue scenarios, directly addressing the primary source of recall degradation observed in this evaluation.
- Expansion of language support to accommodate multilingual and code-switching content, broadening the system's applicability across global platforms.
- Incorporation of video-based visual analysis to detect profanity conveyed through non-verbal cues, graphical overlays, and symbolic representations.
- Evaluation on a larger, more diverse dataset to enable statistically robust performance benchmarking and cross-category generalization testing.

REFERENCES

1. J. Smith and A. Brown, "Automatic profanity detection in online media," Journal of Language Technology, vol. 12, no. 3, pp. 45–52, 2020.
2. Google Cloud, "Speech-to-Text API." [Online]. Available: https://cloud.google.com/speech-to-text
3. M. Jones, "Natural language processing with spaCy," Python Journal, vol. 6, no. 1, pp. 10–15, 2021.
4. K. Lee, "Contextual challenges in speech profanity detection," IEEE Transactions on Affective Computing, vol. 8, no. 2, pp. 114–123, 2019.
5. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
6. N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," IEEE Access, vol. 9, pp. 88364–88376, 2021.
7. P. Malik, P. Bansal, and R. Singh, "Toxic speech detection using BERT and FastText embeddings," in Proc. 5th Int. Conf. Computing Methodologies and Communication (ICCMC), 2021, pp. 1254–1259.
8. G. H. Panchala, V. V. S. Sasank, D. R. H. Adidela, P. Yellamma, K. Ashesh, and C Prasad, "Hate speech and offensive language detection using ML and NLP," in Proc. 4th Int. Conf. Smart Systems and Inventive Technology (ICSSIT 2022), 2022, pp. 1262–1268.
9. A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Language Processing for Social Media, 2017, pp. 1–10.
10. P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.
11. R. Patel, S. Sharma, and A. Gupta, "Detecting toxic comments on social media: An extensive evaluation of machine learning techniques," Journal of Computational Social Science, vol. 8, pp. 20–39, 2024.
12. S. Koenecke et al., "Racial disparities in automated speech recognition," Proc. National Academy of Sciences, vol. 117, no. 14, pp. 7684–7689, Apr. 2020.
13. M. Voss, C. Witschel, and P. Langer, "Measuring the accuracy of automatic speech recognition solutions," ACM Trans. Accessible Computing, vol. 17, no. 1, pp. 1–38, Dec. 2023, doi: 10.1145/3636513.
14. V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, "Comparison of deep learning models and pre-processing techniques for toxic comment classification," Applied Sciences, vol. 10, no. 23, pp. 1–25, 2020.
15. B. Vidgen, T. Thrush, Z. Waseem, and D. Kiela, "Learning from the worst: Dynamically generated datasets to improve online hate detection," in Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP), 2021, pp. 1667–1682.