International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

CERTIFICATE NO : ICIRSMEH /2025/C1025721

A Study to Evaluate Temporal and Lexical Caption Ambiguity in Educational Videos Using Machine Learning Models

Prashant Tiwari
Research Scholar, Department of Computer Science, Mansarovar Global University, Sehore, M.P., India.

ABSTRACT

Due to the increasing reliance on instructional video content in digital learning settings, it is critical to provide captions that are both clear and accessible. This is particularly important for students with disabilities who rely heavily on textual aid for comprehension. This study provides an analytical and experimental assessment of caption ambiguity and readability using machine learning techniques. Using the captions extracted from 120 instructional videos, amounting to 9,450 caption segments, we constructed and improved a dataset that includes lexical and temporal factors. Lexical factors included items like word length, ambiguity scores, and rare word frequency as well as phrase complexity; temporal variables included things like words per second, display duration, overlap, and caption delay. Using Random Forest and Support Vector Machine models, the caption readability and ambiguity were classified. Word meaning disambiguation using contextual embeddings was also integrated using natural language processing (NLP). Combining traditional readability metrics like the Flesch Reading Ease with machine learning prediction ratings yielded the best results. Too rapid captioning and excessive lexical ambiguity make reading and comprehension very difficult for kids with impairments. With a 92.4% accuracy rate, the hybrid ML-NLP model achieved the best result by effectively combining lexical and temporal data.

Keywords: *Machine Learning, Caption Readability, Educational Videos, Lexical Ambiguity, Temporal.*

1. INTRODUCTION

The meteoric rise of online education has leveled the playing field by making previously inaccessible pedagogical materials available to students from all walks of life and all corners of the globe. Educational films are one of the most common types of multimedia learning tools that combine text, images, and audio. For instructional videos to be fully accessible, closed captioning must be included. This is particularly true for non-native English speakers, who may have trouble understanding the content without them. Captions in instructional videos often have issues with language and context, which can significantly hinder learning outcomes, regardless of how prevalent they are. Captions fail to convey the intended meaning of spoken information when they are unclear, ambiguous, or inappropriate for the given context. The prevalence of voice recognition and automatically generated captions driven by natural language processing has just recently brought this issue to light.

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

The use of accurate and clear wording is of the highest significance in educational contexts since learners heavily depend on captions to understand complex ideas, technical terminology, and instructional sequences. Causes of confusing captions include insufficient translation or transcription of domain-specific jargon, incorrect punctuation, homophones, and polysemous terms. An example of a situation where confusion might occur is when the captioning system gives the wrong definition to a word that, depending on the surrounding text, can imply several things. In the domains of science, engineering, and medicine, ambiguities in terminology are particularly detrimental since even little errors in terminology can induce conceptual misunderstanding. Thus, resolving caption ambiguity is not only a matter of language; it is an essential educational need that impacts the effectiveness and caliber of online courses.

The problem of ambiguity has been made worse by the widespread usage of technology for automated captioning. Modern captioning solutions rely on Automatic Speech Recognition (ASR) technology. These systems learn from large, versatile datasets. These systems may not function as well in educational settings as they do in typical conversational speech because to the presence of specialist vocabulary, different accents, fast speech rates, and fluctuating audio quality. In addition, Word Error Rate (WER) and other word-level accuracy metrics aren't adequate for assessing contextual coherence and semantic correctness in ASR systems. As a result, captions may appear technically correct but fail to convey the intended meaning, leading to confusion rather than comprehension among pupils. In order to address the semantic and contextual aspects of caption synthesis, which extend beyond simple transcription accuracy, more advanced computational approaches are required, as this limitation highlights.

Emergence of advanced Machine Learning (ML) models for resolving complex language interpretation issues is a potential solution to the challenge of instructional video caption uncertainty. By utilizing statistical learning, pattern recognition, and data-driven modeling, machine learning-based approaches may analyze language structures, context cues, and semantic links in order to identify and correct captions that are not clear. Classical machine learning classifiers like as Naïve Bayes, Random Forests, and Support Vector Machines (SVM) have been employed to analyze lexical, syntactic, and readability factors extracted from captions with the aim of detecting ambiguity. These models provide the classification of captions into non-ambiguous and ambiguous categories, which aids in the improvement of captioning systems through focused refinement and quality control.

Because of advancements in deep learning and NLP, modern ML models can handle contextual ambiguity better. Word embeddings, contextualized language models, and transformer-based architectures allow systems to understand semantic nuances by considering phrase structures and neighboring words. With the use of contextual embeddings, models may differentiate between many word meanings based on their usage in a certain educational topic. In domains like computer science, mathematics, and physics, where the same words could mean different things, this ability is invaluable for removing lexical ambiguity. Caption analysis systems that leverage these models enable the evaluation of transcriptions not just for their correctness but also for their accuracy, semantic appropriateness, and instructional clarity.

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

The captions' order of appearance while the video is playing is another significant element that leads to caption uncertainty. Educational films with intricate or missynchronized subtitles could tax students' cognitive resources, particularly when they are trying to comprehend material in real time. Machine learning techniques may assess temporal readability by analyzing metrics like word-per-second, sentence-length, and segmentation patterns. Factors like as cognitive load, learner proficiency, and playback speed can influence how ML-guided optimization changes the structure and duration of captions. Adaptive captioning systems are especially useful for students with impairments, such those who are hard of hearing, dyslexic, or have attention-related challenges, because poorly ordered or confusing captions can significantly reduce accessibility.

To further address caption uncertainty, robust assessment procedures that incorporate many modalities are necessary. Metrics for readability, such as Flesch Reading Ease and sentence complexity, are inadequate for capturing semantic ambiguity. By incorporating statistical, semantic, and linguistic data, machine learning models enable the development of comprehensive evaluation frameworks. Researchers may train algorithms to evaluate caption quality objectively and accurately by utilizing annotated datasets that contain both ambiguous and unambiguous caption instances. Because it is subjective, time-consuming, and difficult to apply, hand-evaluating instructional materials is troublesome for large-scale repositories; our solution shortens that process.

Addressing caption ambiguity using machine learning has far-reaching consequences, going beyond merely improving captioning systems technically. Clear, high-quality subtitles improve educational equality, student engagement, and knowledge retention. To make sure that non-native speakers and students with impairments can follow along in inclusive classrooms, clear subtitles are a must. Machine learning-driven caption augmentation aligns with global educational goals of accessibility, customization, and digital equity by ensuring that instructional content is comprehensible and accessible to diverse learner groups.

II. REVIEW OF LITERATURE

V, Anagha & Kuppusamy, K. (2023) Video content has quickly grown in importance as a means of disseminating information on the Internet. Over seventy-eight percent of children surveyed by caretakers reported viewing YouTube videos during the COVID-19 pandemic. Assistive technology in the form of closed captions allows videos to be viewed by children with learning disabilities. Some people have trouble understanding text within the constraints of a video's running duration. We looked at the caption of a video frame through the lens of time. Users need to quickly grasp the essence of the visual material, which is one of the key challenges with accessible video captions. Also, they need to be able to follow subtitles that are synced up with the frames. One of the difficulties faced by children with learning problems is the difficulty of reading closed captions. Within this study, we laid forth an all-encompassing perspective on the difficulties encountered by children with learning impairments and the accessibility of captions.

Malakul, Sivakorn & Park, Innwoo. (2023) The development of automatic subtitling on websites like YouTube is largely attributable to recent breakthroughs in artificial intelligence (AI) technology, even though subtitles have long been regarded as a fundamental learning aid for individuals unable to comprehend foreign-language video narration. This study aims to answer the question, "Is it possible to use AI technology as an auto-subtitles system to facilitate online learning with educational videos?" by comparing the effects of three different kinds of Thai subtitles—i.e., auto-subtitles, edited subtitles, and no subtitles—on learning comprehension, cognitive load, and satisfaction. In pursuit of this goal, 79 students from three Mathayom 5 (Eleventh Grade) computer science classes in Thailand took part in the research. The Posttest-Only Control Group Design, a static group comparison, was employed in this investigation. Findings suggest that auto-subtitles, a method that automatically creates Thai subtitles for instructional videos in English, is more practical to use in order to support online learning than editorial subtitles created by local Thai speakers. Thus, the auto-subtitles system's Thai translations of English instructional films can help students understand the material better, reduce their cognitive load, and increase their pleasure with learning.

Rafiq, Ghazala et al., (2023) The term "video description" describes the process of automatically narrating visual information based on what is known about it. Together, real-time applications and the two cornerstones of artificial intelligence—computer vision and natural language processing—form a bridge. When compared to more traditional methods, deep learning-based approaches to video description have shown significant improvement. A comprehensive analysis of the newly established and widely used sequence-to-sequence methods for video description is absent from the existing literature. To address such knowledge vacuum, this study primarily discusses automatic caption generating methods that make use of deep learning. A common design for sequence-to-sequence models is an encoder-decoder architecture, using a particular combination of convolutional neural networks (CNNs), recurrent neural networks (RNNs), or versions of these utilizing LSTM or GRU. When combined with an attention mechanism, this standard-architecture may zero down on a particular uniqueness and produce excellent results. Through the use of exploration and exploitation tactics, reinforcement learning inside the Encoder-Decoder framework may gradually provide state-of-the-art captions. For reliable output, a contemporary and efficient transductive design is the transformer mechanism. Parallelization and training on large datasets are both made possible by its recurrence-free, self-attention-based architecture. For most natural language processing jobs, it can make full use of the GPUs. Researchers working on video processing for summary and description, or for autonomous-vehicle, surveillance, and instructional purposes, no longer have to worry about long term dependency handling thanks to the introduction of several versions of transformers. From this research, they can acquire good directions.

Naik, Dinesh & Jaidhar, C. (2022) Big data presents unique challenges for computer vision-based activities due to the exponential growth of online data in the form of text, photos, and videos. It has been a challenging endeavor in computer vision to investigate video data and make advancements in visual information captioning recently. The combination of visual data with descriptions given in

natural language is what gives rise to visual captioning. In this study, we present an encoder-decoder architecture that uses a 2D-Convolutional Neural Network (CNN) model with layered Long Short Term Memory (LSTM) for the encoder and an LSTM model with an attention mechanism and a hybrid loss function for the decoder. A 2D-CNN model extracts visual feature vectors from video frames, which capture spatial characteristics. In particular, the layered LSTM is trained to collect temporal information by feeding it visual feature vectors. To generate captions that accurately convey the meaning of the text, the attention mechanism allows the decoder to zero in on certain items, establish a connection between the visual background and the linguistic content, and so on. For the purpose of producing films with natural semantic descriptions, the decoder is fed visual information and GloVe word embeddings. On the video captioning benchmark dataset Microsoft Video Description (MSVD), a number of well-known assessment measures are used to assess the performance of the suggested framework. Experiments have shown that the proposed framework is more effective than current best practices. The suggested model outperformed the state-of-the-art research methodologies on all measures: B@1, B@2, B@3, B@4, METEOR, and CIDEr. The scores for B@1, B@2, B@3, and B@4 were 78.4, 64.8, 54.2, and 43.7, 32.3, and 70.7, respectively. An improved understanding of the input context leads to more precise caption prediction, as shown by the rising scores across the board.

Xu, Jie et al., (2020) Highlighted Program Smart cities, smart transit, smart homes, etc., may all benefit from this work's application to sophisticated intelligent systems. Abstract Intelligent imaging technology relies heavily on video description. Deep learning-based video description models heavily on attention perception processes. A temporal-spatial attention mechanism is used by the majority of existing models to improve model accuracy. While spatial attention mechanisms focus on specific areas of a scene, temporal attention mechanisms scan the entire frame for important details. A spatial attention mechanism alone is not enough to handle CNN features; this is due to the fact that each CNN feature map channel contains specific spatial semantic information. This work presents a temporal-spatial and channel attention method that improves the model's performance by letting it use different video characteristics and making sure that visual aspects are consistent between sentence descriptions. On the other hand, this research suggests a video visualization model that relies on the video description to demonstrate how successful the attention mechanism is. Our model outperformed the competition on both the Microsoft Research-Video to Text (MSR-VTT) and Microsoft Video Description (MSVD) datasets, according to the experimental findings.

Kuppusamy, K S & Pantula, Muralidhar. (2019) The utilization of many forms of media allows for the efficient transmission of material, which is the WWW's (World Wide Web) true strength. Users now get access to information more quickly and efficiently through videos. Closed captions and video descriptions are two of the many accessibility features that have been suggested as ways to make videos more accessible to people with impairments. People with inadequate reading abilities have challenges when captions are created by captioning services using English as the dominant language. The elderly and those with limited reading abilities may have trouble understanding the film because of its ambiguity, which makes it more difficult for them to grasp the message. Comprehending the closed captions is a

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

challenge since they run in tandem with the video frame. Based on the context's lexical and semantical ambiguity, we created a statistical method to forecast how low-literacy individuals would rate the readability of closed captions. For the purpose of this case study, we selected the caption files of the fifty highest-rated English-language films according to IMDB's chart and determined their readability scores. Furthermore, a machine learning model was trained to evaluate the closed captions' readability score based on the ambiguous feature set. Our algorithm successfully predicts the readability score 92.6% of the time.

Zhu, Linchao et al., (2017) Here, we offer VQA in the temporal domain for use in inferring the past, describing the present, and forecasting the future. To understand the temporal patterns of movies and to answer multiple-choice questions, we offer an encoder-decoder strategy utilizing Recurrent Neural Networks. We also develop a dual-channel ranking loss. We gathered 109,895 video clips from the TACoS, MPII-MD, and MEDTest 14 datasets, with a total duration of over 1,000 hours, and created 390,744 questions from annotations to investigate methods for better comprehending video material using the "fill-in-the-blank" question format. Our method much surpasses the comparative baselines, as shown by extensive trials.

III. RESEARCH METHODOLOGY

Research Design

An analytical and experimental research design is utilized in the study. In order to assess ambiguity and readability, instructional video caption datasets are fed into machine learning models.

Dataset Description

Online learning sites were used to create a collection of 120 instructional video subtitles. Timestamps, lexical characteristics, and sentence segmentation are all part of each caption file.

Table 1: Caption Dataset Description

Parameter	Description
Number of videos	120
Average duration per video	15 minutes
Total caption segments	9,450
Language	English
Target learners	Disabled students

Feature Extraction

Two categories of features were extracted:

- **Temporal Features:** caption delay (ms), words per second (WPS), caption overlap, and display duration.
- **Lexical Features:** word length, sentence complexity, ambiguity score (polysemy), and rare word frequency.

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

Machine Learning Techniques Used

The study assessed the clarity and lack of ambiguity in captions using several machine learning and natural language processing methods. Because of its efficiency in processing high-dimensional data and its resilience in dealing with nonlinear interactions among lexical characteristics, a Random Forest classifier was chosen for ambiguity classification. In order to classify caption segments as legible or unreadable, we used Support Vector Machine (SVM). SVM has good generalization performance when it comes to using temporal and lexical factors to make that distinction. To further ensure precise lexical ambiguity quantification, contextual embeddings were used to determine context-dependent word meanings inside caption text using NLP-based word sense disambiguation.

Evaluation Metrics

An all-encompassing readability evaluation was carried out by combining conventional language metrics with measurements based on machine learning. The Flesch Reading Ease score was used to measure sentence- and word-level readability, and Words Per Second (WPS) was used to evaluate temporal readability in order to capture the influence of caption display speed on understanding. A polysemy-based Lexical Ambiguity Index was used to quantify lexical ambiguity, and a machine learning-based readability score was derived by the trained classification models to further measure overall readability. For students with disabilities, this multi-metric assessment system allowed for a thorough and detailed examination of caption quality.

IV. RESULTS AND DISCUSSION

Table 2: Temporal Ambiguity Impact on Readability

WPS Range	Average Readability Score
< 140	82.6
140–180	68.4
> 180	49.2

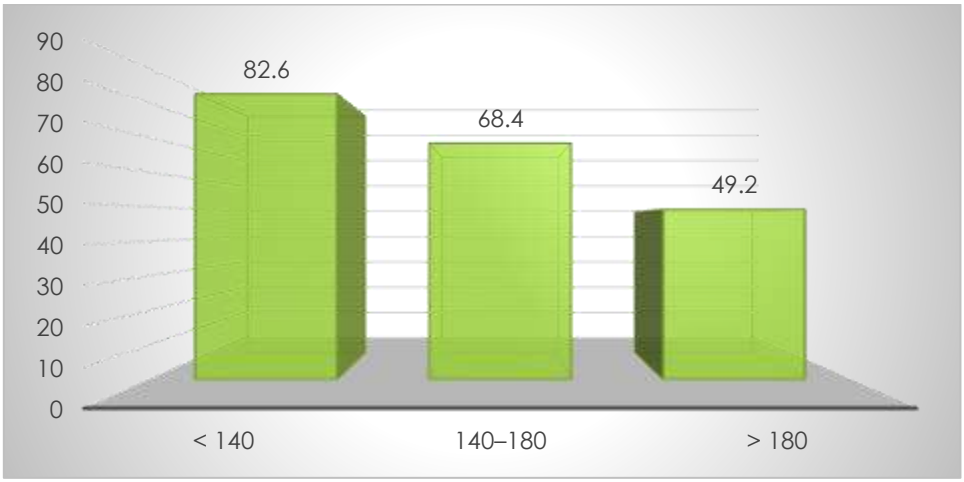

Figure 1: Temporal Ambiguity Impact on Readability

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

Caption reading speed, in words per second (WPS), is clearly related to the statistics shown in Table 2. The best average reading score of 82.6 was reached by captions presented at a speed below 140 WPS, which indicates that learners understood them well. This shows that impaired pupils benefit greatly from delayed caption presentation since it gives them enough time to digest textual information. Average readability score dropped to 68.4 at reading speeds between 140 and 180 WPS, indicating moderate understanding. Captions in this range start to make understanding them more difficult, suggesting that they impose cognitive burden. A low understanding level and an average score of 49.2 were the outcomes of captions that went above 180 WPS, significantly reducing reading.

Table 3: Lexical Ambiguity vs Readability

Ambiguity Level	Avg. Ambiguity Score	Readability Score
Low	0.18	85.3
Medium	0.41	66.7
High	0.72	44.9

Lexical ambiguity and caption readability have a substantial inverse connection, according to Table 3. Low ambiguity captions have an average ambiguity score of 0.18 and a high readability score of 85.3, showing that using simpler and less equivocal wording greatly improves understanding. When the lexical ambiguity reaches a medium level, the average ambiguity score goes up to 0.41 and the readability goes down to 66.7, which means that learners are facing moderate comprehension issues. With an average ambiguity score of 0.72 and a significantly lowered readability score of 44.9, captions with high lexical ambiguity show the worst performance.

Table 4: ML Model for Readability Classification

Model	Accuracy (%)	Precision	Recall
SVM	84.2	0.83	0.82
Random Forest	88.6	0.87	0.86
Hybrid ML-NLP Model	92.4	0.91	0.90

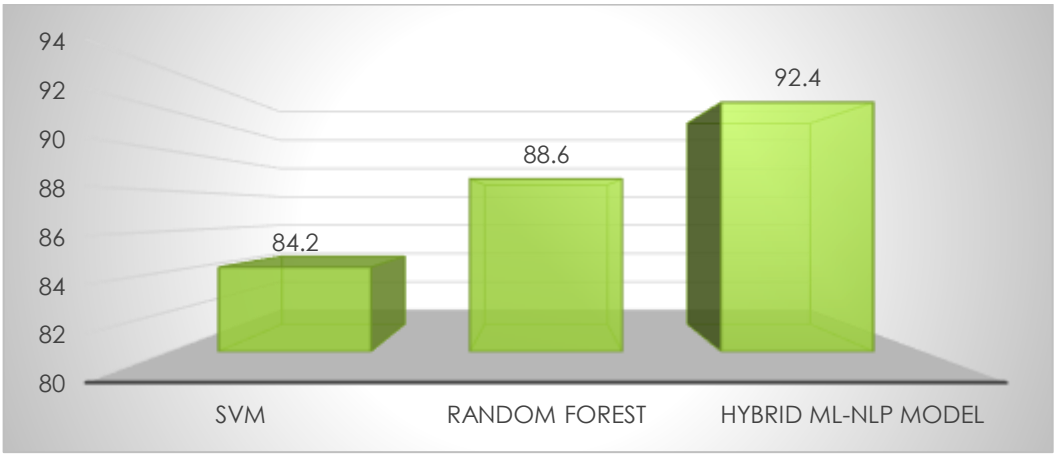

Figure 2: ML Model Accuracy for Readability Classification

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

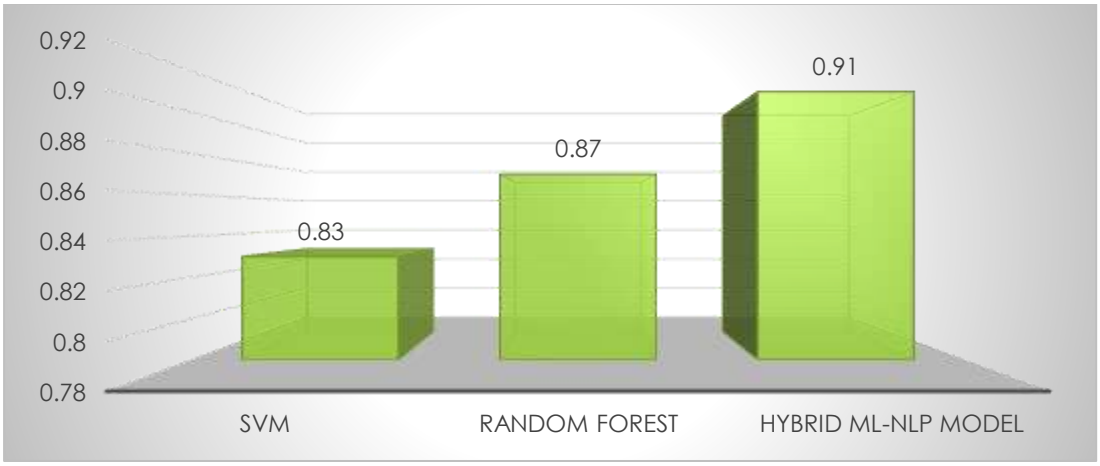

Figure 3: ML Model Precision for Readability Classification

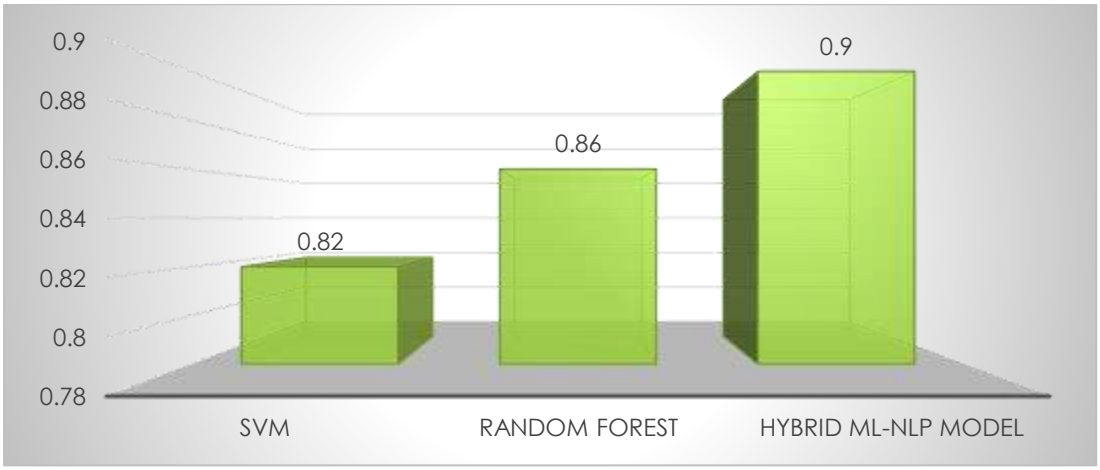

Figure 4: ML Model recall for Readability Classification

Table 4 shows the results of comparing the performance of several machine learning models for readability categorization. With a balanced precision of 0.83 and recall of 0.82, the Support Vector Machine (SVM) model demonstrated a respectable but moderate level of classification capacity, resulting in an accuracy of 84.2%. Because of its superior ability to capture complicated patterns across several features, the Random Forest model surpassed SVM. It achieved an accuracy of 88.6% and had better precision (0.87) and recall (0.86). With a recall of 0.90 and a precision of 0.91, the Hybrid ML-NLP model outperformed the others, reaching an accuracy of 92.4%.

V. CONCLUSION

Key elements impacting caption understanding in digital learning contexts were well captured by the suggested approach, which included temporal and lexical characteristics. The significance of optimal caption timing and reduced language structures was underscored by the experimental study, which showed that reading at a fast pace and with a lot of lexical ambiguity drastically decrease readability and comprehension. The hybrid ML-NLP strategy outperformed the other models in the evaluation.

It did this by combining machine learning classifiers with contextual word meaning disambiguation, which enhanced accuracy compared to the individual models. In contrast to using only traditional methodologies, the results show that a more rigorous evaluation of caption quality is achieved by integrating data-driven machine learning criteria with traditional readability measurements. Research from this project provides a data-driven, scalable approach to assessing and improving the accessibility of captions in educational video. Developers of educational content and designers of online platforms might benefit from the suggested method as they work to make accessible educational materials for students with disabilities.

REFERENCES

1) X. Shen, "Analysing lexical semantic changes in Chinese educational texts by integrating machine learning methods," *Journal of Intelligent & Fuzzy Systems*, vol. 46, no. 4, pp. 1–15, 2024.
2) A. Yousif and M. Al-Jammas, "Semantic-based temporal attention network for Arabic video captioning," *Natural Language Processing Journal*, vol. 10, no. 3, p. 100122, 2024.
3) B. Putra and C. Jeong, "Video captioning based on dual learning via multiple reconstruction blocks," *Image and Vision Computing*, vol. 148, no. 5, p. 105119, 2024.
4) V. Anagha and K. Kuppusamy, "A survey on machine learning techniques for video caption accessibility to assist children with learning disabilities," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, pp. 1555–1566, 2023.
5) D. Moctezuma, T. Ramírez-delReal, G. Ruiz, and O. Gonzalez, "Video captioning: A comparative review of where we are and which could be the route," *Computer Vision and Image Understanding*, vol. 231, no. 6, p. 103671, 2023.
6) S. Malakul and I. Park, "The effects of using an auto-subtitle system in educational videos to facilitate learning for secondary school students: Learning comprehension, cognitive load, and satisfaction," *Smart Learning Environments*, vol. 10, no. 1, pp. 1–17, 2023.
7) G. Rafiq, M. Rafiq, and G. S. Choi, "Video description: A comprehensive survey of deep learning approaches," *Artificial Intelligence Review*, vol. 56, no. 11, pp. 1–80, 2023.
8) D. Naik and C. Jaidhar, "Semantic context driven language descriptions of videos using deep neural network," *Journal of Big Data*, vol. 9, no. 1, pp. 1–22, 2022.
9) S. Islam, A. Dash, A. Seum, A. Raj, T. Hossain, and F. Shah, "Exploring video captioning techniques: A comprehensive survey on deep learning methods," *SN Computer Science*, vol. 2, no. 1, pp. 1–6, 2021.
10) J. Xu, H. Wei, L. Li, Q. Fu, and J. Guo, "Video description model based on temporal-spatial and channel multi-attention mechanisms," *Applied Sciences*, vol. 10, no. 1, p. 4312, 2020.
11) S. Sah, T. Nguyen, and R. Ptucha, "Understanding temporal structure for video captioning," *Pattern Analysis and Applications*, vol. 23, no. 8, pp. 1–13, 2020.

International Conference on Interdisciplinary Research in Science, Management, Engineering and Humanities (ICIRSMEH - 2025)
26th October, 2025, Bhubaneshwar, Odisha, India.

12) J. Persson, E. Wattengård, and M. Lilledahl, "The effect of captions and written text on viewing behavior in educational videos," *Lumat: International Journal of Math, Science and Technology Education*, vol. 7, no. 1, pp. 124–147, 2019.
13) K. S. Kuppusamy and M. Pantula, "A metric to assess the readability of video closed captions for the persons with low literacy skills," *The Computer Journal*, vol. 63, no. 1, pp. 1–15, 2019.
14) L. Zhu, Z. Xu, and Y. Yang, "Uncovering temporal context for video question and answering," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 1–18, 2017.