



**INAOE**

# **Attention-Based Multimodal Learning for Hate Speech Detection in Videos for Mexican Spanish**

by

**Itzel Tlelo Coyotecatl**

Doctoral Advisor:

**Dr. Hugo Jair Escalante Balderas, INAOE**

**Technical Report No. CCC-24-003**

January, 2024

©Coordinación de Ciencias Computacionales  
Instituto Nacional de Astrofísica, Óptica y Electrónica

Luis Enrique Erro 1  
Sta. Ma. Tonantzintla,  
72840, Puebla, México.



# Attention-Based Multimodal Learning for Hate Speech Detection in Videos for Mexican Spanish

## Abstract

*Nowadays, social media and internet platforms allow us to consume, generate, and share content in the easiest possible ways. This content is usually regulated by conduct rules that enforce a healthy coexistence and interaction among users. However, there are some cases where this content has harmful intentions directed at a person or a particular group, either because of race, gender, sexual orientation, nationality, or religion, among others. Such behavior is associated with the term hate speech.*

*To regulate and identify the presence of hate speech content, works in computer science have been proposing automatic methods that use features, extracted from available modalities (e.g., text, images, etc.), to allow the detection of hate speech. Despite the good performance in the task, these methods still lack of techniques that leverage the relationship between modalities when analyzing complex content like images or video. Additionally, a lack of research on content other than English is remarkable. Therefore, this research work is focused on the hate speech detection task for video content in Mexican Spanish using multimodal information. Because most of the state-of-the-art works are oriented to English-based content we propose the construction of a Mexican Spanish video dataset that will later serve as a reference in the expansion of this research field.*

*On the other hand, to leverage the use of multimodal information on the hate speech detection task we propose to work on an automatic method that will be capable of learning multimodal embeddings that aim to be used by state-of-the-art models to address the task. Lastly, at the end of this research work, we expect to achieve: 1) A novel automatic method that successfully learns multimodal representations that enhance the performance of state-of-the-art models to address the hate speech detection task, 2) A Mexican Spanish video dataset that expands the available resources for the hate speech detection task. By achieving these expectations we remark on the significance not only of the hate speech detection task, considering the social implications, but also on the expansion of resource availability for the task and the exploration of techniques that leverage the multimodal representation for complex content.*

**Keywords**— *Hate Speech Detection, Vision & Language, Multimodal Information*

# Contents

<b>Abstract</b>	<b>1</b>
<b>1 Introduction</b>	<b>4</b>
1.1 Motivation . . . . .	4
1.2 Justification . . . . .	4
1.3 Problem Statement . . . . .	6
1.4 Research Questions . . . . .	6
1.5 Hypothesis . . . . .	7
1.6 Objectives . . . . .	7
1.7 Expected Contributions . . . . .	8
<b>2 Background</b>	<b>9</b>
2.1 Feature Extraction Techniques . . . . .	9
2.1.1 Text Modality . . . . .	9
2.1.2 Visual Modality . . . . .	10
2.1.3 Audio Modality . . . . .	11
2.2 Multimodal fusion Techniques . . . . .	12
2.2.1 Multimodal information . . . . .	12
2.2.2 Multimodal Representation . . . . .	13
2.2.3 Multimodal Fusion . . . . .	13
2.3 Attention Mechanism (Transformer) . . . . .	14
2.4 Classification Methods . . . . .	16
2.4.1 Classic Machine Learning . . . . .	16
2.4.2 Deep Learning . . . . .	17
2.4.3 Transfer Learning . . . . .	17
2.5 Performance Indicators . . . . .	18
<b>3 Related work</b>	<b>20</b>
3.1 Hate Speech Detection . . . . .	20
3.1.1 Overall approaches . . . . .	20
3.1.2 Shared tasks on Hate Speech datasets . . . . .	21
3.2 Research Proposal vs. Similar Works . . . . .	21
3.3 Transformers for Video Classification . . . . .	23
<b>4 Methodology</b>	<b>25</b>
4.1 Mexican Spanish video dataset creation . . . . .	25
4.1.1 Collection of videos from social media . . . . .	26
4.1.2 Annotation of the videos . . . . .	27
4.1.3 Formalization of the dataset . . . . .	27

4.2	Multimodal method proposal . . . . .	27
4.2.1	Extraction of multimodal features . . . . .	27
4.2.2	Fusion of multimodal features . . . . .	28
4.2.3	Use of the multimodal representation . . . . .	31
4.2.4	Evaluation of the multimodal method . . . . .	31
<b>5</b>	<b>Preliminary Results</b>	<b>33</b>
5.1	Towards a Dataset for Hate Speech Detection in Videos . . . . .	33
5.1.1	Collecting and processing videos from YouTube. . . . .	33
5.1.2	Annotating the dataset . . . . .	34
5.1.3	Analyzing the current dataset . . . . .	37
5.1.4	Formalizing our dataset . . . . .	38
5.2	Hate Speech Detection Task using Multimodal Information . . . . .	39
5.2.1	Extraction of multimodal features . . . . .	39
5.3	Preliminary Conclusions . . . . .	40
<b>6</b>	<b>Final Remarks</b>	<b>42</b>
6.1	Future Work . . . . .	42
6.2	Work Plan . . . . .	43
	<b>Bibliography</b>	<b>44</b>

# 1 Introduction

## 1.1 Motivation

Nowadays, there are approximately 4.76 billion social media users around the world<sup>1</sup>. Consequently, the use of social media has become a daily part of our routine, we consume, generate, and share various content. Despite the efforts of social media platforms to monitor and regulate the content, there are still cases where certain users promote the presence of offensive and/or abusive content. The *Hate Speech* (HS) phenomenon can be described as “any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factors” [58]. In order to detect and regulate the presence of HS in social media, efforts in the computer science field have been increasing by developing automatic methods focused on the HS detection task. These methods aim to take advantage of features extracted from the published available content (e.g., text, images, audio, etc.) [46]. Thus, it became usual to use these features to train classifiers that determine if the analyzed content should be classified as hate speech content or not.

Despite the good performance of existent automatic methods for the hate speech detection task, it is noteworthy a lack of research for languages other than English as well as methods aimed at analyzing complex content (e.g. videos). To begin with, the lack of available resources different from English for the task is noticeable. Considering the creation and use of resources different from English suggests challenges like the scarcity of resources availability and language-based implications related to language structure, context, and/or culture. As far as we know, no previous works address Mexican Spanish video content. Thus, the richness of idioms, and building a dataset from scratch are challenges that must also be considered. On the other hand, most social media involves complex content that includes images and/or videos. Then the analysis and use of two or more modalities should be considered to represent that content by taking advantage of the information that each modality could provide. The adaptation of automatic methods to a multimodal approach that is capable of representing the interaction between modalities and using that information to perform competitively on the hate speech detection task is a noticeable necessity. Certainly, this implies other derived challenges like modality fusion and representation, temporality alignment, or resource availability.

## 1.2 Justification

The hate speech detection task has gained interest in recent years. In the computer science field, works related to the Natural Language Processing (NLP) area have addressed the task in different ways considering the variations in the definition of hate speech as a

---

<sup>1</sup>Global Social Media Statistics. <https://datareportal.com/social-media-users>

concept. Particularly because the hate speech definition has been evolving in order to deal with phenomena related to abusive messages, insults, offensive language, and vulgar language, among others. The problem of detecting hate speech is considered highly relevant because of its direct influence on the interaction between communities. Due to this reason, monitoring social media content with capable and trustworthy automatic methods is a must.

In the computer science field, automatic methods are required to be adapted to use different kinds of information to accomplish the hate speech detection task. Traditionally, automatic hate speech detection models have been taking advantage of features extracted from social media content. These features can be extracted from text, images, audio, and other modalities. Once the features are extracted an acceptable representation of the content is required to serve as input to train the computational models. Then, many state-of-the-art techniques highlight the use of unimodal and/or multimodal approaches that face the hate speech detection task in multiple ways. It is common to address the task as a binary classification problem, where the main goal for the proposed automatic methods is to determine whether the provided content is predicted as hate speech or not. Additionally, there also exists work that further research on detecting the category of hate speech content. Subsequently, the hate speech detection task, oriented to social media content, faces relevant challenges related to feature extraction techniques for the available modalities, the fusion of multimodal information techniques as well as the data representation in a shared latent space, and the proposal of novel good-performing automatic methods that leverage the use of the available modalities of the provided content.

It is noteworthy to mention that existing state-of-the-art works for the task present some limitations. The identified limitations include a predominant focus on English-based content, a limitation on the quantity and quality of the resources to be working on, and an inclination to work on text modality or bi-modal approaches for text-visual cues extracted from images content. Thus, considering the mentioned limitations there are two main points we aim to address with this work.

First, we identified that most of the work in the hate speech detection task focuses on content that uses modalities like text and images [46], as a consequence, there is a lack of analysis for complex social media content (e.g. videos). It is remarkable the aspiration to develop models aimed at analyzing this complex media content, by integrating multiple modalities. Particularly, our intention is to focus on the leverage of fusion and representation of the multimodal available information. We expect that the integration of modalities like text, visual, and audio may provide us with the essential information that represents in a reliable way the content. We suspect there would be videos where certain modalities would be more relevant. Thus, by considering the possible integration of text, visual, and audio we expect to balance the contribution of each one into the general reliable multimodal representation.

Second, we identified that the majority of the literature reports results addressing the task for English language-based resources [43]. Then, a lack of experimentation in other lan-

guages is notable not only by the language itself but also by the limited available datasets in languages different from English. Therefore, we consider the importance of creating datasets in languages other than English. Because Spanish is becoming increasingly globalized since it is the official language of 21 countries and is spoken by over 474.7 million native speakers<sup>2</sup>, it is expected that the creation of a Spanish language-based resource will later serve as a reference in the expansion of this field of research. Particularly, we intend to create and refine a Mexican Spanish video dataset that as far as we know, no other work has done it before.

In order to achieve the mentioned points, our proposal considers the design of a novel automatic method aimed at learning multimodal embeddings that generate an acceptable representation of the interaction of the available modalities into a shared latent space. We expect to use the multimodal representation in state-of-the-art models to enhance the performance of the hate speech detection task considering its application on video content.

### 1.3 Problem Statement

This research proposes to address the problem of the *hate speech detection task* in video content through the development and implementation of computational methods using multimodal information. Additionally, considering that most of the work is oriented toward English language content, and few of them explore languages other than English, we propose the building of a video dataset focused on the Mexican Spanish language.

We outstand some of the main challenges we will have to overcome: 1) there is a lack of literature focused on hate speech detection from videos; 2) there is a lack of video resources, particularly there are no resources for Mexican Spanish, and language itself implies changes related to the context, interpretation of idioms, and cultural implications; 3) working with video content modalities (e.g., audio, visual, text) may be challenging because of the differences in data formats leading to non-trivial techniques to integrate the multiple multimodal features considering factors like temporality alignment, and modeling of context and interaction between modalities.

### 1.4 Research Questions

Attending to the previously stated problem statement the main research questions are listed below.

- How feasible is it to construct a valuable Mexican Spanish video dataset for the hate speech detection task?
- How feasible is it to develop an effective hate speech detection method for video content considering the Mexican Spanish context?

---

<sup>2</sup>The most spoken languages in the world. <https://www.berlitz.com/blog/most-spoken-languages-world>

- What approaches should be adopted to learn and use effective multimodal embeddings to enhance performance on the hate speech detection task in videos when used with attention-based models?
- How does contemplating multiple language modalities (e.g., English captioning, Spanish transcriptions) in addition to multiple available modalities (e.g., visual, audio) will benefit our method’s ability to detect hate speech content?

## 1.5 Hypothesis

Our main hypothesis states that a novel multimodal method oriented to perform the hate speech detection task in Mexican Spanish videos, that successfully leverages multimodal information by homogeneously integrating the multiple modalities into a valuable embedding representation processed by an attention-based model, may obtain remarkable results that outperform actual state-of-the-art approaches.

Specifically, our multimodal method is planned to leverage the generation and use of multimodal representations that effectively capture modality interactions by simultaneously processing the multiple modalities and fusing them into a multimodal embedding. Then applying an attention-based model to the multimodal embedding representation will allow us to capture valuable information in order to enhance the performance of the hate speech detection task in videos.

## 1.6 Objectives

The main objective is to design and develop an automatic method that leverages the use of multimodal information by learning multimodal representations to achieve competitive performance that outperforms the state-of-the-art approaches, on the hate speech detection task in Mexican Spanish video content. The specific objectives are listed below:

- To create a Mexican Spanish video content-based dataset for the hate speech detection task to serve as a reference in the evaluation of the proposed method and the expansion of this field of research.
- To explore and assess various feature extraction and fusion techniques to leverage multiple modalities into a successful representation of video content.
- To design an effective method for generating multimodal embeddings that capture and represent meaningful interactions and information between multiple modalities.
- To use and evaluate the generated multimodal representation into state-of-the-art methods to analyze the performance opportunity improvements.
- To evaluate the performance of the novel proposed automatic method to compare it to state-of-the-art indicators.



## 1.7 Expected Contributions

This work is expected to focus on the research and experimentation with automatic methods that use multimodal information for the hate speech detection task-oriented to Mexican Spanish video content. It is intended to focus on social media video content by leveraging an acceptable representation of the features extracted from multiple modalities (e.g., text, images, audio) to improve the detection of hate speech. Also, the creation of a Mexican Spanish video dataset aims to expand the benchmarking opportunities of this research field. To sum up, the main goal of this research is to analyze competitive automatic methods in the state of the art in order to propose a novel automatic method that uses multimodal information to successfully perform the hate speech detection task. Therefore, the main contributions of this research work are:

- A novel multimodal method oriented to the hate speech detection task in Mexican Spanish video content.
  - The multimodal method will generate an effective multimodal embedding representation by simultaneously processing the available modalities and fusing them into an embedding that leverages the information and interactions between them.
  - The multimodal method will process the multimodal embedding representations by applying an attention-based model to enhance the performance when detecting the possible presence of hate speech content.
- A new social media video dataset that broadens research field benchmarking opportunities for the Spanish language, emphasizing Mexican Spanish.

## 2 Background

Traditionally the hate speech detection task has been addressed by a common approach that involves the extraction of features that serve as input for a classification method in order to obtain a final prediction that indicates if the analyzed content presents hate speech or not [46]. This section will provide an overview of the most commonly used feature extraction techniques (Subsection 2.1), multimodal fusion techniques (Subsection 2.2), attention mechanism (Subsection 2.3), classification methods (Subsection 2.4) and evaluation indicators (Subsection 2.5) considered for this research work.

### 2.1 Feature Extraction Techniques

In computational approaches, feature extraction refers to the process of selecting and transforming raw data into a set of meaningful features that can be used as inputs for computational models. Feature extraction is an important step in many computational approaches, as it enables the model to effectively learn patterns and relationships in the data. A brief overview of the most common feature extraction approaches is summarized in the sections below for the most used modalities (e.g. text, visual, and audio).

#### 2.1.1 Text Modality

Common feature extraction techniques in Natural Language Processing (NLP) for text modality may vary from simple features, word generalization, sentiment analysis, lexical resources, and knowledge-based features. It will depend on the essence of the task to be tackled. In the case of working with video content, captions, and transcriptions are treated as text modalities [16].

**Word generalization.** This technique allows us to understand the meaning of words beyond their specific context. In NLP, word embeddings [9] is a popular technique used for word generalization. It can be defined as a dense vector representation of words that capture semantic and syntactic relationships between them by mapping each word to a dense vector in a high-dimensional space, where the distance between vectors reflects the similarity of the words. The representations are typically learned through unsupervised learning algorithms such as Word2Vec [38] and GloVe [42], which process large text corpora to learn a distributed representation of the words. The use of this technique provides us with the advantage of capturing similarities and differences in words meaning that are not explicitly expressed in the text. But it also has to be considered that the representation may be limited by the vocabulary and/or context the models were trained on.

**Lexical Resources.** In NLP, depending on the task, the presence of certain words can be considered as a feature. Then, some authors incorporate lexical resources into their methods. Lexical resources are collections of words or phrases with associated linguistic information (e.g., part-of-speech (POS) tags, semantic labels, or sentiment scores). Some

works incorporate dictionaries and/or lexicons specially compiled for a task to identify particular content related to the presence, frequency, and weights of content words or phrases related to the task [46, 16]. Lexical resources prove a source of knowledge about language that can be used to improve the accuracy of NLP models. However, the quality and coverage may vary depending on the source and the domain, so they may need to be adapted or extended for specific applications.

### 2.1.2 Visual Modality

In the computer vision field, when processing visual modalities (e.g., images, frames of a video), two main techniques are highlighted, Convolutional Neural Networks (CNN or ConvNet) and Transformers [39].

- CNN models take images as input, process them into several layers (e.g., convolutional layer, pooling layer, fully-connected layer), and produce as an output a final result based on the previously learned features.
- Transformer models take images as input and involve two main components, a linear projection of the image and an encoder. The encoder is composed of several multi-layer perception neural network models and a self-attention mechanism to capture relevant information that represents the input.

Despite its popularity and the good reported results for numerous tasks, CNN may require large datasets to be trained on, high computational requirements, and may be difficult to interpret. Then, to facilitate the interpretation, Transformers present an advantage when capturing global information because of their self-attention mechanism. This allows it to receive entities without context and deliver output with contextual information by prioritizing the relevant content and learning representations from various perspectives because of its multi-head attention element. Because of their differences, both techniques have been adapted depending on the task and in some cases pre-trained alternatives or hybrid approaches have been explored too.

The selection of the technique for visual modality feature extraction may vary depending on the task but in general, when dealing with video, spatial and temporal features are considered.

- Spatial feature extraction implies obtaining a description of a relative spatial area of an object and its spatial semantic relationships with other objects by working on still video frames [53]. This spatial information allows to differentiate static and dynamic contexts. Examples of classical methods include corner detectors, edge detectors, Harris detector, Hessian detector, or sampling methods. More complex recent methods involve the use of deep neural networks. They usually are an extension of classical image data vision techniques. This means working with multiple frames over time to learn the spatial features.

- Temporal feature extraction implies recognizing the action or movement of an object from motion in the form of an optical flow [6]. Optical flow may be described as a motion pattern of an object in staked video frames [53]. Classical descriptors include dense trajectory, histogram of optical flow (HOF), and its variation histogram of oriented optical flow (HOOF). For analysis of long-term optical flow features, graphical models (e.g., bayesian networks, hidden Markov models, conditional random fields) are used. More recent models include the use of deep neural networks (e.g., convolutional neural networks, and recurrent neural networks). Additionally, networks like FlowNet2 [27], LiteFlowNet [25], and EV-FlowNet [65] have reported obtaining good accuracy results.

Because video content involves both spatial and temporal features, systems like the two-stream approach and multi-model hybrid approach have been developed [53]. While the first one focuses on the combination of procedures on spatiotemporal features from stacked optical flows and still frames, the second one is focused on abstract features such as long-term motion cues. As an example of a well-known visual feature extractor, Inflated 3D ConvNet (I3D) is a popular model, for video classification and action recognition. **I3D** [10] model is defined as a Two-Stream Inflated 3D ConvNet (I3D) that is based on 2D ConvNet inflation. The 3D expansion of filters and pooling kernels of very deep image classification ConvNets make it possible to learn spatiotemporal feature extractors from video. Typically this model takes video clips as input in the format of RGB frames and produces as output a set of feature vectors that represent spatio-temporal information (e.g. RGB, and flow vectors).

### 2.1.3 Audio Modality

Audio feature extraction is done from an audio signal. The process considers the removal of unwanted information (noise) and also considers both, time and frequency features. These features comprehend an analysis of the audio signal in its original form. Then, a spectral representation of the signal is done in the frequency domain [37]. Classic machine-learning approaches may include techniques like Amplitude Envelope, Zero-Crossing Rate, Root Mean Square Energy, Spectral Centroid, Band Energy Ratio, and Spectral Bandwidth. On the other hand, most recent deep learning approaches work with spectrograms or Mel Frequency Cepstral Coefficients (MFCC) by extracting patterns.

The most common feature extraction techniques for speech recognition [31, 51] include:

- **Spectrogram.** Defined as a two-dimensional visual depiction of an input image. Signals can be plotted considering time and frequency.
  - **mel-spectrogram.** Is a technique where frequencies are converted into a logarithmic scale (mel scale).

- **MFCC**. Also known as Mel Frequency Cepstral Coefficients, this technique chops the speech signal into frames and then applies Fast Fourier Transform (FFT) on each frame to obtain the power spectrum.
- **Linear Prediction Coefficients (LPC)**. This technique provides vectors of speech based on a frame analysis of the speech signals. The process includes the spectral flattening of the digitized input signal, frame blocking of the acoustic signal, removal of signal discontinuities using a hamming window, apply of the autocorrelation method, and LPC parameter set conversion.

Depending on the task pre-trained models for audio feature extraction can be used. One well-known accessible model for audio feature extraction is VGGish. **VGGish**<sup>3</sup> is a variation of the VGG [24] model. This model takes as input mel spectrograms, computed by the short-time Fourier transform (STFT) of the audio signal mapped onto the mel scale. This allows the extraction of semantically meaningful embedding features. Then Principal Component Analysis (PCA) is used to extract normalized features to produce as output a 128-dimension feature vector [33]. The model is trained on AudioSet [19] dataset which contains over 2 million human-labeled 10s YouTube video soundtracks with more than 600 audio event classes.

## 2.2 Multimodal fusion Techniques

Working with multiple modalities implies the consideration of fusing them into a common representation that captures the interaction between modalities. In the following sections, we will describe the multimodal information concept (Subsection 2.2.1), techniques for multimodal representation (Subsection 2.2.2), and ways to fuse the modalities (Subsection 2.2.3).

### 2.2.1 Multimodal information

Nowadays, the implementation and use of multimodal approaches arises because of the complexity of the analyzed content [11]. Multimodal information implies the use of two or more available modalities where a *modality* refers to how something is expressed or perceived. The main objective will be to work on heterogeneous (diverse) and interconnected data. The main challenges [35, 7] associated with multimodal data integration and fusion may consider the acceptable representation of the data interactions, the alignment of the data by identifying and modeling its connections, the combination of knowledge (reasoning), the generation of raw modalities that reflect their original interactions, the transfer of knowledge between modalities and the quantification (a study to better understand the learning process, interconnection and heterogeneity of modalities).

---

<sup>3</sup><https://github.com/tensorflow/models/tree/505f9bf352d35700bf2596f7d9ce908881f81a07/research/audioset/vggish>

## 2.2.2 Multimodal Representation

Representation is considered a relevant point when talking about summarizing the available multimodal data. Both joint representations and coordinated representations are considered techniques to accomplish it.

**Multimodal joint representations** may include simple concatenation, element-wise multiplication or summation, and multilayer perceptron as possible techniques [35]. This implies an explicit modeling of the  $n$  modality interactions. The result representations of processing  $n$  modalities are reduced in most cases to one aiming to project the modalities into a joint space [64].

**Coordinated multimodal representations** are aimed to implicitly learn the representations from the multiple available modalities by implementing more complex methods. Also, instead of projecting into a shared space, separated representations are learned for each modality and then they are coordinated through a constraint [64]. The result representations of processing  $n$  modalities are still the same quantity but these new representations integrate the information between modalities. Some examples include similarity models and structured models. The similarity models aim to minimize the distance between modalities in the coordinated space by using neural networks. Structured models enforce additional constraints based on the application (e.g., hashing, cross-modal retrieval, and image captioning).

## 2.2.3 Multimodal Fusion

When working with multiple modalities a representation that involves the best of each one is required [62]. There are two main ways to process the input modalities into a unified result: model-agnostic and model-based fusion.

**Model-agnostic** involves working on early, late, or hybrid stages. *Early fusion* is one of the most commonly used techniques because of the easy implementation (concatenation of features) done before feeding the classifier. Despite the ease of its implementation, it can also end up as a high-dimensional result. *Late fusion* requires multiple training stages that involve the training of unimodal predictors and then the implementation of the multimodal fusion one. A loss of some low-level interactions should be considered. Additionally, the *hybrid fusion* involves the combination of both already mentioned mechanisms.

**Model-based** fusion involves techniques like kernel methods, graphical models, or neural networks. *Kernel learning* considers using kernels for each modality and learn which one are important for the classification. They were considered as an extension of kernels of support vector machines (SVM), where the kernels are seen as similarity functions. These models are quite popular because of their implementation flexibility but also it has to be considered that their reliance on training data could lead to slow inference and large memory footprint.

Existing *deep learning* models for multimodal data fusion traditionally use a deep model to capture specific features per modality (e.g., neural networks like RNNs and LSTMs).

This allows to transform the raw representations into a high-abstraction representation that can be visualized into a global space [18]. Also, when using these models some considerations should be addressed, this includes being aware of computing cost and time-consuming, the uncertainty of collected data, and semantic data relationships.

### 2.3 Attention Mechanism (Transformer)

Transformers were first introduced in 2017 in the paper "Attention is All you Need" [56] (Figure 1) approaching the machine translation task. They were described as a network architecture based solely on attention mechanisms. Its architecture involves two main blocks, the encoder and decoder architecture, as it is shown in Figure 1. The encoder block is in charge of taking an input and generating a fixed-length contextualized output vector. Then, that output is processed by the decoder to obtain an output sequence (e.g., the contextualized vectors return to words in the case of language translation). But not all the tasks require both, encoder and decoder components, the architecture can be adapted to capture the necessary information.

The popularity of transformers is related to their ability to process whole sequences at once, which allows parallelizing operations [47] and capturing relevant global information about the input. The core component of the transformer architecture is the self-attention mechanism that allows capturing relevant dependencies between elements in the input sequence. For example, if the input sequence is a sentence of words, after applying the self-attention the words with more relevance will be represented with higher weight values.

$$Att(Q, K, V) = softmax \left( \frac{QK^T}{\sqrt{d_k}} \right) V \tag{1}$$

In the overall process, attention is calculated by Equation 1, and it can be described as mapping a query and a set of key-value pairs to an output where the query, keys, values, and output are all vectors. Then, the output is computed as a weighted sum of values. Because the attention function is computed for a set of queries, dot-product attention is done for the queries, keys, and values packed into matrices. Additionally, to attend to information from different representation subspaces at different positions instead of performing a single attention function a multi-head attention approach is applied (Figure 2).

Despite transformers being at first designed for language translation, their popularity has been increasing because of their generalization and adaptability to other tasks including not only text-related ones but also vision (e.g., ViT [13]), audio or multi-modality approaches [1]. However, it is important to consider that the main limitations are related to the quadratic time complexity when working with long sequences, and the lack of inductive biases that imply the possibility of slowing down learning unless large quantities of data are used [47]. In consequence, the use of pre-trained versions derived from a vanilla transformer or other similar architectures may be an alternative for particular tasks or specialized datasets. The work [3] provides an extensive catalog of available transformer

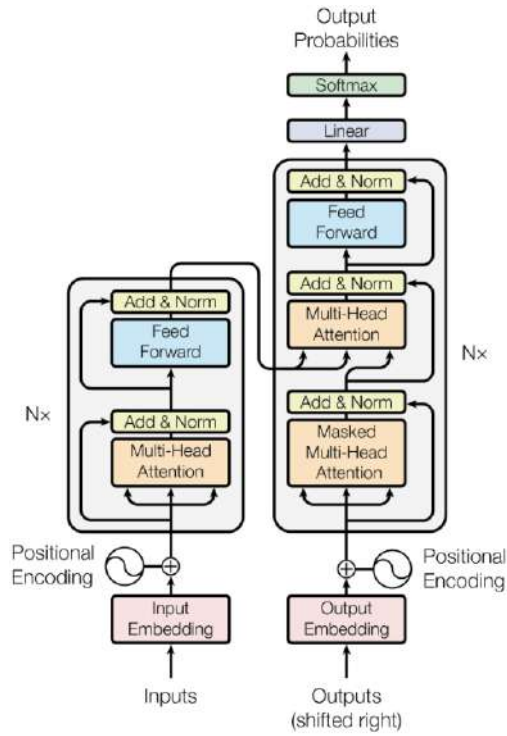


Figure 1: Transformer architecture figure from [56]

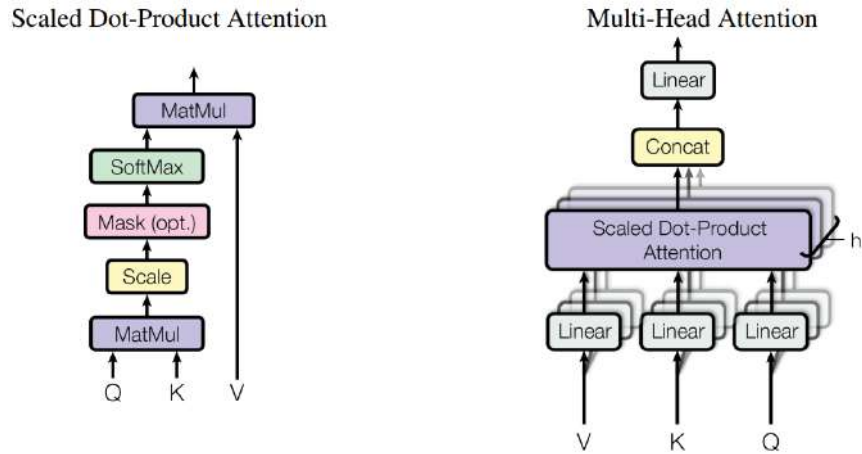


Figure 2: Transformer attention mechanism figure from [56]

variations depending on properties like pertaining architecture, pertaining task, compression application, or the number of parameters.



## 2.4 Classification Methods

Classification methods have been evolving significantly over the years, driven by advances in machine learning and artificial intelligence research. Through the years we have seen methods that manually define a set of rules for data classification (Rule-Based methods), methods that were able to handle larger and more complex datasets (Classic Machine Learning methods), methods that allow handling high-dimensional data and achieved impressive results in a wide range of applications (Neural Networks), methods that used multiple layers of interconnected nodes to extract complex features from the input data resulting on a great performance and becoming the state-of-the-art standard in many areas (Deep Learning), methods that use pre-trained models as a starting point for new tasks rather than training a new model from scratch (Transfer learning) and methods that aim to be more transparent and interpretable (Explainable AI methods). The following sections show a brief description of classic machine learning methods, deep learning methods, and transfer learning.

### 2.4.1 Classic Machine Learning

Classic machine learning methods are a set of techniques and algorithms used to train models on input data in order to learn patterns and relationships between the input features and output labels. Once the model is trained on a dataset it can be used to make predictions on new data. Some of the most used classic machine learning methods include:

- **Decision tree** models that learn a hierarchical set of rules based on the input features to predict an output label.
- **Naive Bayes** models rely on probability by estimating the probability of each considered class given the input features and selecting the class with the highest probability as the predicted label.
- **Support vector machine (SVM)** models that learn hyperplanes to separate the input data into different classes [40, Ch. 7].
- **Random forest** ensemble method that combines multiple decision trees to improve the accuracy and robustness of the model.
- **K-nearest neighbors (KNN)** model that predicts the output label based on the k-nearest data points in the training set.

While these methods have been used in a wide range of NLP tasks their limited ability to handle complex data and capture nuanced relationships, between the input features and output labels, resulted in the necessity of developing more advanced methods such as neural networks and deep learning.

### 2.4.2 Deep Learning

As a subfield of machine learning deep learning has gained significant popularity and success due to its ability to solve complex problems related to various areas (e.g., image and speech recognition, and natural language processing). Deep Learning (DL) core relies on artificial neural networks (NN) which were modeled and inspired by the structure and function of the human brain. It consisted of layers of interconnected nodes where each one performs a simple computation on its input and passes its output to the next layer. By training the NN on a large dataset, the weights of the connections between the neurons are adjusted to minimize the error between the predicted output and the actual output. Popular DL architectures include:

- **Convolutional Neural Networks (CNN)** use convolutional layers to extract features from the input, commonly images, and pooling layers to reduce the dimensionality of the features.
- **Recurrent Neural Networks (RNN)** use recurrent connections to maintain a memory of previous inputs and use this information to make predictions. This architecture is commonly used for sequential data.
- **Generative Adversarial Networks (GAN)** involve two networks working together to generate realistic data. One network generates fake data and the other tries to distinguish between fake and real data.
- **Transformer Networks** use attention mechanisms to allow the model to selectively focus on certain parts of the input, allowing for more effective processing of long sequences of data [3].

Despite the impressive performance of DL methods one of the main issues is their requirement of large amounts of training data and computational resources, which can make them difficult to apply in some tasks. Additionally, the models can be difficult to interpret, which can make it challenging to understand why certain predictions were made.

### 2.4.3 Transfer Learning

Transfer learning is a technique that uses pre-trained models [17] as a starting point for training a new model on a related task. There are two main types of transfer learning: fine-tuning and feature extraction.

**Fine-tuning** involves taking a pre-training model and training all or some of its layers on a new task, typically with a lower learning rate to avoid destroying the learned features.

**Feature extraction** involves using the pre-trained model is used as a feature extractor to generate a set of high-level features that are relevant to feed into a new set of layers that are trained specifically for the new task. Some popular pre-trained models for transfer

learning include VGG, ResNet, and Inception for image recognition, BERT and GPT-2 for NLP, and WaveNet for speech recognition.

Transfer learning can significantly reduce the amount of labeled training data required to achieve high performance on new tasks by reusing features previously learned by pre-trained models on related tasks. Additionally, the overfitting problem can be overcome because of the no dependence on too specialized training data. However, transfer learning may be considered cautiously because of the limited applicability, possible data biases, required computational resources and training time, as well as the limited control over the pre-trained used model.

## 2.5 Performance Indicators

In order to evaluate the performance in the classification made by a computational model, standard indicators are required. The task itself requires analyzing the classification of the samples. Then, a **confusion matrix**, as a tool, is considered to identify the number of correct and incorrect previously classified samples. By comparing the resulting label classification generated by the computational model to the golden truth there are four possible categories that the sample could be placed on. These categories are described below.

- **True Positive (TP)**. It is when the golden truth label of the sample originally belongs to the positive class and the predicted label generated by the computational model match with the positive class.
- **True Negative (TN)**. It is when the golden truth label of the sample originally belongs to the negative class and the predicted label generated by the computational model match with the negative class.
- **False Positive (FP)**. It is when the golden truth label of the sample originally belongs to the negative class but the predicted label generated by the computational model ends up in the positive class.
- **False Negative (FN)**. It is when the golden truth label of the sample originally belongs to the positive class but the predicted label generated by the computational model ends up in the negative class.

All the categories (TP, TN, FP, FN) belong to each one of the cells indicated on the confusion matrix shown in Table 1. It is expected that the generated classification achieves high results for the TP and the TN cases.

To evaluate the performance of the classification, indicators like Precision (P), Recall (R), and F1-score are calculated by using the resulting values in the confusion matrix. The corresponding equations for each indicator are presented below.

		Predicted class	
		Positive	Negative
Golden truth Class	Positive	TP	FN
	Negative	FP	TN

Table 1: Confusion Matrix

$$P = \frac{TP}{TP + FP} \quad (2)$$

$$R = \frac{TP}{TP + FN} \quad (3)$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (4)$$

### 3 Related work

This section will present details about the state of the art for the hate speech detection task (Subsection 3.1), similar works to our research proposal (Subsection 3.2) and an analysis of video transformer existent related work (Subsection 3.3).

#### 3.1 Hate Speech Detection

##### 3.1.1 Overall approaches

The hate speech detection task has been evolving in recent years. Several state-of-the-art approaches have been proposed [46] and can be broadly categorized into three categories: unimodal approaches, bimodal approaches, and multimodal approaches.

Unimodal approaches typically focus on using either text, audio, or visual information to detect hate speech but mostly text is used [16]. The most common unimodal approach for hate speech detection in videos considers text modality to work with [2, 28, 60]. This implies working with transcriptions of the audio from the video or comments associated with the video. Feature extraction techniques such as bag-of-words, n-grams, and word embeddings are commonly used to represent the text content. Then, a variety of machine learning and deep learning models are trained with the features to obtain a final prediction for hate speech detection.

Bimodal approaches including text-audio [44] or text-images [21] modalities for video content not only extract the correspondent modality features but also apply a fusion technique to generate a complete representation of the involved modalities to work with. Then, features such as pitch, intensity, and facial expressions are used to represent the audio or visual information that can be contemplated. The resultant representation embedding vector is used as input to a computational model (e.g., classical, deep learning, transfer learning approaches) to generate the hate speech predictions for the content. These approaches [44, 21] have shown promising results, but it is still required to adapt them in order to identify a variety of forms of hate speech that may rely on the combination of different or more modalities.

On the other hand, multimodal approaches aim to combine information from multiple modalities to improve the accuracy of hate speech detection, as well as to analyze complex content. As far as we know, there is no work focused on the hate speech detection task that combines more than two modalities. Considering that the current most effective multimodal approaches use deep learning models to learn representations of the data and make classifications for the tasks [11]. Because hate speech as a phenomenon could be expressed in a variety of ways we aim to leverage our multimodal approach by integrating modalities like visual, audio, and text. We expect each modality to provide us with the essential information that once integrated into a multimodal representation allows us to comprise the video content.

### 3.1.2 Shared tasks on Hate Speech datasets

The International Workshop on Semantic Evaluation (SemEval)<sup>4</sup> is one of the main series of international Natural Language Processing (NLP) workshops that aims to promote the development of NLP technologies through shared tasks and evaluations. Therefore, SemEval provides a platform for researchers to collaborate and compare their approaches to various NLP tasks. Also, the workshop has led to the development of several benchmark datasets and state-of-the-art models in various subfields of NLP like sentiment analysis, entity recognition, information extraction, and others.

	Task	Dataset	Modality	Language
SemEval-2019 [8]	Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter	Twitter	Text	English Spanish
SemEval-2020 [63]	Task 12: Multilingual Offensive Language Identification in Social Media	Twitter	Text	Arabic Danish English Greek Turkish
SemEval-2020 [49]	Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!	Google Images	Image-Text	English
SemEval-2022 [15]	Task 5: Multimedia Automatic Misogyny Identification	Twitter, Reddit	Image-Text	English

Table 2: SemEval tasks through different years editions.

Table 2 shows some of the proposed tasks, related to the hate speech phenomenon, in the different editions of the SemEval workshop through recent years. The tasks were at first delimited to address the detection of hate speech content aimed at immigrants and women, as well as offensive expressions, in text posts. Then, text was the main modality to work with for the proposed unimodal approaches. More recent editions incorporate the analysis of image-text posts (e.g., memes) searching for phenomenons like misogyny or metaphors. This suggests the necessity to monitor more complex content (e.g., videos) through the research and adaptation of multimodal approaches. Additionally, we emphasize the realization that most of the hate speech detection tasks and works [43, 12], have been oriented to analyze English language content. In consequence, an actual concern is the lack of both resources and models adapted to work in other languages.

### 3.2 Research Proposal vs. Similar Works

Table 3 presents a summary of some of the most similar works considered to compare and contrast our research proposal approach. The main common point is the attempt to detect hate speech-related content on video sources.

Anand et al.[4] developed a framework for customized video filtering on YouTube that can detect and filter content based on specific user preferences. They generated their own YouTube video dataset, explored various combinations of features from text and image

<sup>4</sup>SemEval website <https://semeval.github.io/>

Work	Year	Dataset	Dataset size	Language	Modality	Features	Approach	Computational Model
Anand et al. [4]	2019	YouTube	20k	English	Text Images	Text embeddings Meta-information Image embeddings	Deep Learning	LSTM with Hierarchical Attention Network & CNN
Shang et al. [48]	2019	YouTube	1k	English	Text	Topological and Semantic features Linguistic features Meta-information	Classic Deep Learning	XGBoost, Logistic Regression, SVM, Random Forest MLP, ULMFiT, SCNE, RNN-GRU
Alcántara et al. [2]	2020	YouTube	400	Portuguese	Text	N-grams (words & characters)  Word embeddings (single & each word)	Classic - - Deep Learning - Transfer learning	Naive Bayes, Logistic Regression, SVM, Random Forest - - CNN, LSTM - BERT, ALBERT
Wu & Bhandary [60]	2020	YouTube	300	English	Text	TF-IDF, uni-grams, bi-grams (words)	Classic Deep Learning	Naive Bayes, Random Forrest, SVM RNN
Urbano et al. [28]	2021	TikTok	1k	Filipino/Tagalog	Text	Text embeddings	Transfer learning	BERT
Rana & Jha [44]	2022	Twitter YouTube	1k	English	Text Audio	Text embeddings Audio embeddings	Deep Learning	BERT, ALBERT
<b>Our Proposal</b>	<b>2026</b>	<b>YouTube</b>	<b>8k</b>	<b>Spanish</b>	<b>Text Visual Audio</b>	<b>Multimodal embeddings</b>	<b>Deep Learning</b>	<b>Attention mechanisms</b>

Table 3: Research proposal similar works for the hate speech detection task in videos.

modalities and applied some classic machine learning methods and deep learning architectures to evaluate their approach. Shang et al. [48] created an end-to-end supervised learning approach to classify hatred-vulnerable videos from hateful and hatred-free ones by exploring the structure and semantic features of comments, transcription, title, and description from the Youtube analyzed videos. Alcántara et al. [2] provided an analysis of how classic, deep learning, and transfer learning algorithms perform by using different representations of features, both meta-data and text transcriptions, extracted from videos. They created their own YouTube video dataset (OffVidPT-2 and OffVidPT-3) for the evaluation. Wu & Bhandary [60] applied classic and deep learning approaches to classify hate speech related content (normal, hateful, racist, sexist) by processing video transcriptions of their own created video dataset. Urbano et al. [28] proposed the use of BERT for automatic hate speech detection by using speech transcription-based features. The evaluation was made on their own created TikTok dataset. Finally, Rana & Jha [44] proposed a multimodal deep learning framework that combines auditory features that represent emotion and semantic features to detect hateful content. Their approach was evaluated on their own Hate Speech Detection Video Dataset (HSDVD) composed of Twitter and YouTube videos and focused on hate speech targeted towards gender, sexual orientation, autistic minorities, Muslims, Jews, Sikhs, Latinos, Native Americans, and Asians.

Our research proposal aims to create a novel multimodal method oriented to analyzing video content with a particular interest in Mexican Spanish. The dataset is going to be created by us and planned to be released to later serve as a point of reference for the extension of the hate speech detection available resources. For the processing of video content, we will incorporate the analysis of multiple modalities like text, visual, and audio. Aiming to generate a multimodal representation that captures the interaction and information between the modalities by fusing them and representing them in a common shared space. Considering the multimodal representation an attention-based model will be designed to take the multimodal representation as input and leverage the identified relevant informa-

tion that will enhance the prediction of the possible presence or absence of hate speech content.

To sum up, because there is no standard benchmark video dataset most of the recent works focused on multimodal approaches choose to create their own datasets. Most of them are based on the English language, which suggests that analysis and work for other languages are required to expand the hate speech detection task. About the used modalities for video feature extraction, text, and images (frames from the video) are the most used on an unimodal or bimodal approach but because of the complexity of video content new ways of modality usage require to be explored. Finally, state-of-the-art standard relies on the implementation of deep learning models and particularly on transformers. Therefore, novel ways to fusion modality information are required to adapt the models in order to achieve good performance on the hate speech detection task.

### 3.3 Transformers for Video Classification

Nowadays, the analysis of video data has been raising interest in the research community. Because video classification may be a generic task and diverse datasets may be available, there is a tendency for works to address the task by relying on transformer-based architectures [47]. These approaches aim to take advantage of multimodal features by adapting the traditional transformer architecture [56] for video classification.

Model	Year	Modality	Token	Backbone
ViViT [5]	2021	V	P	Linear Layer
CBT [52]	2019	V T	C	S3D [61]
STiCA [41]	2021	V A	C	R(2+1)D-18 [55], RN-9 [23]
PE [32]	2021	V A	C	SlowFast [14], RN-50 [23]
VATT [1]	2021	V A T	P	Linear Layer
VATLM [66]	2022	V A T	F	Multipath transformer [50]
<b>Our Proposal</b>	<b>2026</b>	<b>V A T</b>		

Table 4: Examples of relevant transformer-based models for video classification [47]. Modality corresponds to V=Visual, A=Audio, and T=Text tasks. Input token corresponds to P=Patch, C=Clip, and F=Frame.

Table 4 shows some examples of well-known transformer-based models for video classification. Video Vision Transformer (ViViT) [5] is a pure transformer-based model aimed at capturing spatiotemporal correlation between frames by processing a series of patches from the video. Contrastive Bidirectional Transformer (CBT) [52] is a BERT model variant that processes continuous features as input and learns long-term temporal representations by adopting a stacked transformer architecture, generalizing its training objective to maximize the mutual information between masked signals, and bidirectional context, via contrastive loss. Space-Time Crop & Attend (STiCA) [41] model captures rich temporal signal by first



introducing a feature crop method to simulate spatial augmentations from video and then processing the features by a lightweight temporal transformer. Parameter Efficient multi-modal transformer (PE) [32] model consists on an end-to-end trainable bidirectional transformer architecture that learns contextualized audio-visual representations of long videos that also integrates a novel parameter reduction technique that shares parts of weight parameters across layers and transformers. Video-Audio-Text Transformer (VATT) [1] is aimed at learning multimodal representations from input raw signals by using convolution-free transformer architectures. Visual-Audio-Text Language Model (VATLM) [66] is a cross-modal representation learning framework that uses simple modality-dependent modules to preprocess each modality input and then a multipath transformer to unify the information.

Being aware of works focused on the development of multimodal transformer-based models for classification tasks (Table 4) and considering the complexity and dimensionality of working with videos. Some main concerns about transformer-based models oriented to video content are related to complex architectures and computational costs. Then, video transformer-based architectures may adopt two main approaches when aiming for an efficient design [47]. The first one decomposes all-to-all attention into several smaller attention operations while the second one progressively aggregates information across layers. Despite the aim for efficient transformers in the mentioned approaches, transformers may be biased to separately focus on different features that may affect the content information representation. Thus, to not directly rely on the transformer architecture we plan to work first on the input multimodal representation to provide the transformer with an effective multimodal representation that later, with the attention mechanism application, may leverage the performance on the video classification task.

## 4 Methodology

In order to accomplish the previously stated objectives (Subsection 1.6), towards the proposal of an automatic method that uses multimodal information to address the hate speech detection task in video content, the methodology is presented in detail in the following sections (Figure 3). First, details for the construction from scratch of the proposed Mexican Spanish language-based video dataset will be described (Subsection 4.1). Then, the considered steps to propose a novel multimodal automatic method will be presented (Subsection 4.2).

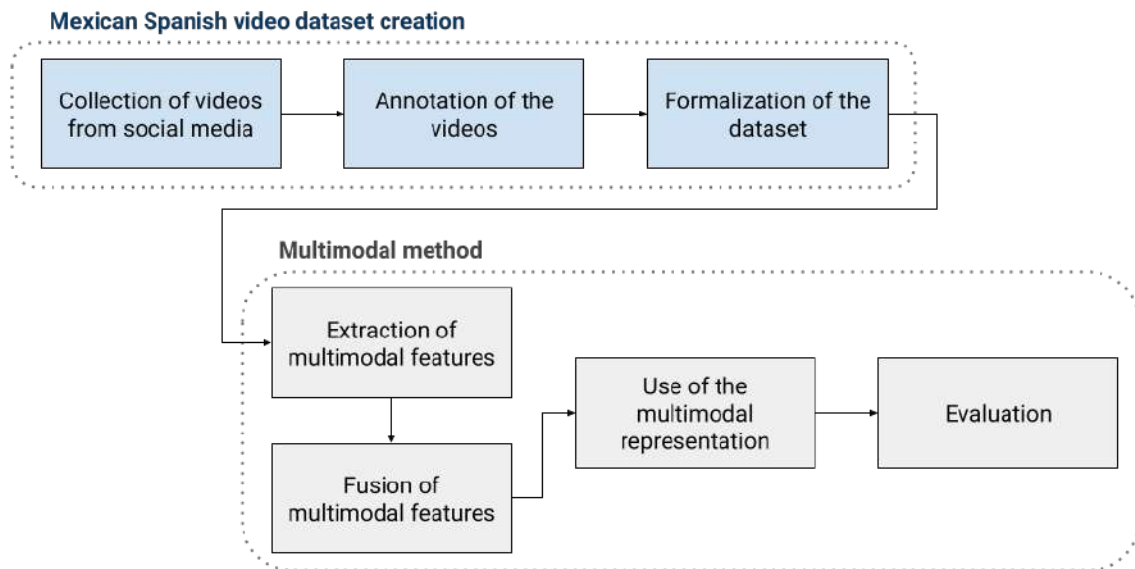


Figure 3: General overview of the methodology.

### 4.1 Mexican Spanish video dataset creation

The creation from scratch of a Mexican Spanish-based video data set for the hate speech detection task is aimed to expand and contribute to the research field in the task. The process involves stages like the collection of videos from social media (Section 4.1.1), the annotation of the videos (Section 4.1.2), and the formalization of the video dataset (Section 4.1.3) as the main steps to follow up. A general overview of the involved stages is shown in Figure 4.

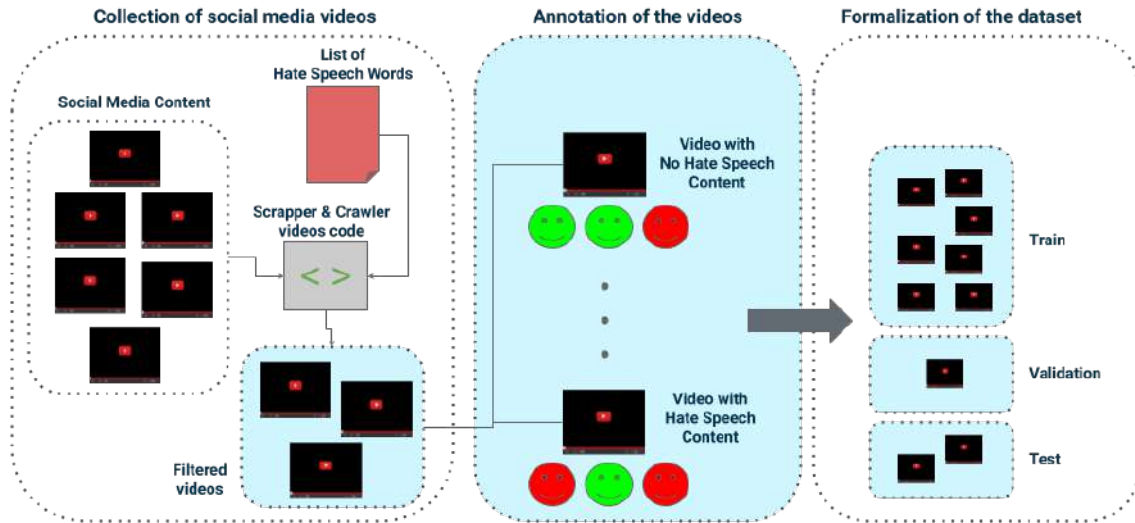


Figure 4: Overview of the stages for the creation from scratch of a Mexican Spanish video dataset for the hate speech detection task.

#### 4.1.1 Collection of videos from social media

Videos considered for the Mexican Spanish video dataset will be collected directly from social media platforms (e.g., YouTube, Twitter, Reddit). The collection of videos will be done by considering the following points:

- A video scraper and crawler will be implemented to download publicly available videos from social media platforms, particularly YouTube.
- A list of Mexican Spanish-related hate speech words/terms will be defined to be used as seeds to filter videos with possible hate speech content presence from social media. The list will be constructed by consulting and referencing works that address the hate speech detection task by constructing their own corpus [43], campaigns that promote to tackle the hate speech phenomenon<sup>5</sup> and other resources.
- A manual revision of a first batch of downloaded videos will be done to select the ones that could be used as seeds to retrieve more videos related to that content.
- A manual revision and navigation on the social media platform will be done to identify possibly hate speech content-related videos (e.g., YouTube publicly available playlists and channels would be explored).

<sup>5</sup>An example of a resource that invites us to reflect on gender equality by exemplifying sexist phrases that we should avoid: <https://www.gob.mx/conavim/articulos/frases-sexistas-que-hombres-y-mujeres-debemos-dejar-de-decir-para-promover-la-igualdad-de-genero?idiom=es>

- The collected videos will be processed by segmenting each one into 1-minute length scenes.

#### **4.1.2 Annotation of the videos**

A manual labeling of the videos will be done considering the already downloaded and processed video batches. In order to do the annotation the following points may be considered:

- An annotation guideline will be created to define the main concepts that annotators will have to be aware of. These concepts will emphasize the hate speech definition, some sub-types of hate speech (e.g., misogyny, violence, discrimination) definitions as well as information that may serve to identify the possible presence of hate speech content.
- An annotation platform will be developed to facilitate the assigned video batches to annotators.
- An analysis of the annotation results will be done by considering annotation agreement indicators to determine the final label for each video.

#### **4.1.3 Formalization of the dataset**

A training, validation, and test partition will be done on the annotated videos to formalize the Mexican Spanish video dataset for the hate speech detection task.

### **4.2 Multimodal method proposal**

In order to address the hate speech detection task by designing and developing a novel multimodal method, that leverages the representation of multiple modalities of information, four stages will be considered. The first stage will address the extraction of each feature from the video source (Section 4.2.1). Then, a fusion of the extracted features will be done by applying a fusion method (Section 4.2.2) to generate an acceptable representation of the video content. This representation will be used as input for an attention-based model (Section 4.2.3) to enhance the hate speech detection task performance. Finally, the fourth stage will evaluate the method’s performance classification by using state-of-the-art standard indicators (Section 4.2.4). Figure 5 shows an overview of the considered stages for the multimodal proposed method.

#### **4.2.1 Extraction of multimodal features**

Because we will work with multiple modalities in this stage we aim to extract features from each representation to later fuse them into a unified representation. The following points will be considered for this stage:

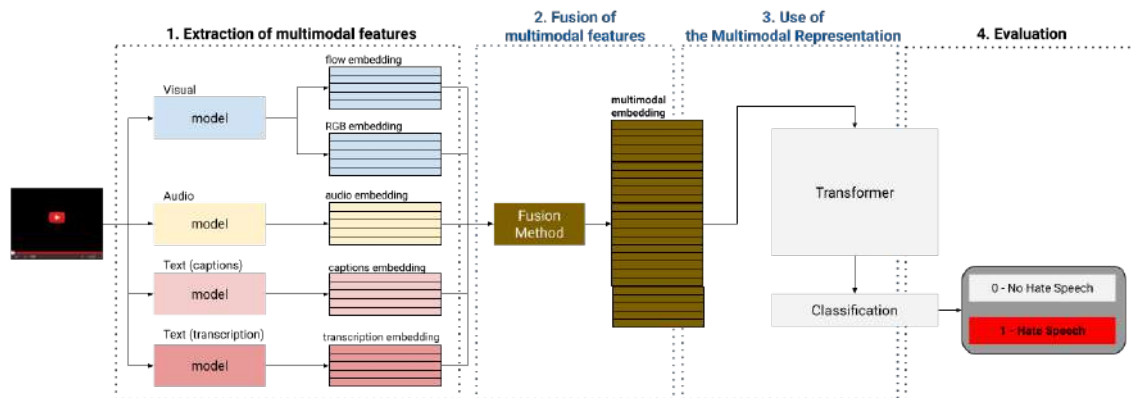


Figure 5: Overview of the considered stages to propose a novel automatic method for the hate speech detection task.

- A literature review on feature extraction methods will be done to determine the use of viable methods according to the available modalities in video content.
  - For text modality methods that involve captioning and automatic speech recognition to generate transcriptions will be explored. Then, Natural Language Processing (NLP) techniques (e.g., word embeddings) may be used to obtain embedding representations of the words.
  - For visual modality, computer vision techniques for image/video processing will be applied to analyze the content (e.g., attention mechanisms like ViViT [5]).
  - For audio modality, audio processing methods will be considered to capture information about background sounds or voice quality (e.g., attention mechanisms like VATT [1]).
- Selected feature extraction models will be applied to the video scenes from our dataset to extract each modality representation (e.g., text embeddings, audio embeddings, visual embeddings).

#### 4.2.2 Fusion of multimodal features

Once we have obtained each modality feature representation we will fuse them into a unified representation that captures interactions between the modalities. Then, a literature review of fusion modality methods will be done to establish a checkpoint from where variations will be explored by designing and proposing techniques that leverage the use of multimodal features to generate a valuable multimodal embedding representation of the video content.

Particularly, our planned solution will require us to emphasize our research in techniques like local linear embeddings, learning of representations, and the use of external kernels to

fusion modalities. To follow up, we describe the possible alternatives we consider to explore in order to design our **multimodal method that generates multimodal embeddings** from given modalities.

**Locally Linear Embedding for multimodal representations.** Working with multiple modalities implies working with a variety of features. Then, integrating them into a shared space should help to represent the common and most relevant information better. Also, high-dimensional data may be a concern when searching for compact representations. Then, dimensionality reduction techniques have been proposed. Particularly, Locally Linear Embedding (LLE) [45] technique focuses on mapping its input into a single global coordinate system of lower dimensionality by calculating the nearest neighbors for each data point, calculating a weight matrix with weights that best reconstruct each data point from its neighbors, and computing vectors that best reconstruct the points into the new lower dimensional space (Figure 6).

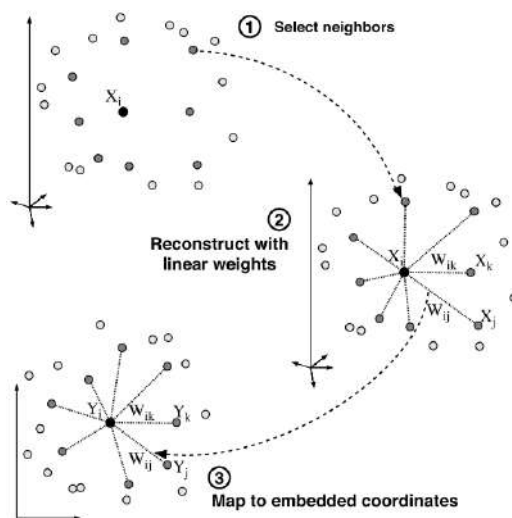


Figure 6: General steps for Locally Linear Embedding technique [45].

Thus, our idea is to adapt the technique to handle as input each modality from video content (e.g., visual, audio, text) and generate as output a new embedding vector of low dimensionality with the context of each provided modality. To achieve this we might explore: capturing relationships within each modality separately and then integrating them into a final embedding (Figure 7) by applying concatenation as the simplest way of fusion or a linear combination that considers assigning weights to each modality; or include a cross-modal step when generating the joint representation for the multimodal embeddings. Either the variation when comparing the original data to the generated multimodal embeddings will serve as a reference to verify how well the multimodal embeddings represent the original information.

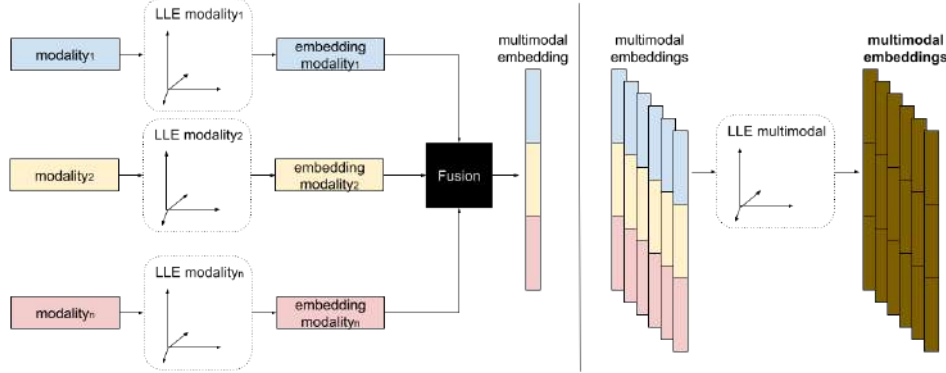


Figure 7: General overview of possible multimodal embeddings generated with LLE.

**Learning of the multimodal representations.** Working with multiple modalities requires capturing their interactions and connections when generating a representation that involves all the provided modalities. Because we are going to work with video multiple modalities (e.g., audio, visual, text) we will need a way to integrate them into a multimodal embedding representation. To achieve it we are planning on exploring abstract fusion representation by learning some weights that allow us to capture the necessary information of each modality and integrate it into a shared representation. The general idea is to capture unimodal representations and then fuse them to learn a joint representation [34]. This fusion could be done by a fusion layer and then a network will be intended to learn the multimodal joint representation (Figure 8).

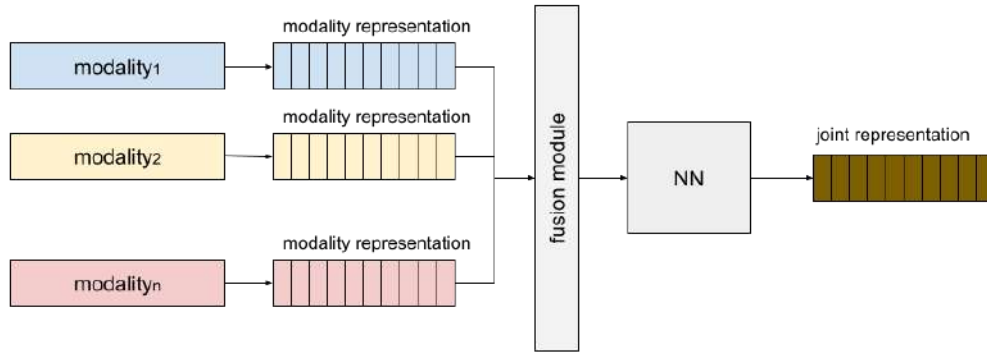


Figure 8: General overview of possible multimodal embeddings generated by learning the multimodal representation.

**Using kernel to fusion modalities.** Working with multiple modalities requires considering different data formats and information. Then, the use of kernels may help to

capture interactions within each modality at first, and then fuse the founded characteristics to generate the multimodal embedding representation of the content (Figure 9). It is expected that the use of multiple modality kernels help us to maintain the heterogeneity of each modality [22] for later handling in a flexible way the final multimodal representation.

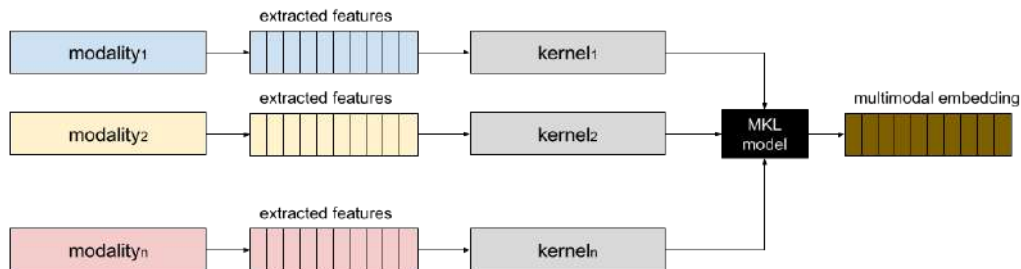


Figure 9: General overview of possible multimodal embeddings generated by using multiple Kernels.

### 4.2.3 Use of the multimodal representation

The generated multimodal representation will be used as input for state-of-the-art methods, mainly attention mechanism models, to address the hate speech detection task. As far as we know, transformers are the standard of the state of the art. Therefore, transformer architectures will be explored, and/or necessary modifications will be done to the multimodal representation or to the automatic method to leverage the available information to achieve competitive performance on the task. At first, a pre-trained transformer model would be finetuned with our multimodal embeddings. Later, an exploration to enhance and adapt the attention mechanism for multimodal information will be done.

### 4.2.4 Evaluation of the multimodal method

Our multimodal method will be evaluated particularly on the hate speech detection task considering our proposed Mexican Spanish video dataset by calculating state-of-the-art standard indicators. These indicators will correspond to a binary classification evaluation (e.g., precision, recall, f1-score, and accuracy). Additionally, a multi-label classification evaluation will be done considering some sub-categories of hate speech (e.g., misogyny, discrimination, violence).

On the other hand, resource datasets for video classification tasks and hate speech oriented to video datasets would be explored and evaluated using both our multimodal method and some state-of-the-art methods as baselines to compare with. These methods may include deep learning approaches and mostly attention mechanisms oriented to use



a multimodal approach. For example, CBT [52] with visual-text modalities, STiCA [41] with visual-audio modalities and VATT [1] with visual-audio-text modalities.

## 5 Preliminary Results

This section will describe the preliminary results obtained for the built-up of our proposed Mexican Spanish video dataset for the hate speech detection task (Section 5.1) and advances about the use of our dataset for the task involving multiple modalities from the video content (Section 5.2). Then, in Section 5.3 we will present some preliminary conclusions.

### 5.1 Towards a Dataset for Hate Speech Detection in Videos

The hate speech detection task has been gaining interest in recent years. But, the lack of experimentation in languages other than English, as well as, in modalities different from text is notable in the availability of datasets. In this regard, we present the first version of our Mexican Spanish Video Dataset, which is built by collecting YouTube videos with possible hate speech content. Following the previously presented methodology in Section 4.1 we detail the actual state of our video dataset.

#### 5.1.1 Collecting and processing videos from YouTube.

By implementing a video scraper and video downloader scripts using the YouTube API we already collected publicly available videos following the next detailed process:

1. Using hate speech-related word lists as seeds we filtered around 800 videos from YouTube. We built up two lists<sup>6</sup>: one of 29 words based on hate speech terms retrieved from the HateBase<sup>7</sup> platform, and the other of 56 words based on the "*Guía de lenguaje incluyente y no sexista*"<sup>8</sup> from the *Secretaría de Relaciones Exteriores*. Both lists consider Mexican Spanish terms related to possible hate speech content presence. Approximately,  $n=10$  videos were retrieved for each word. Examples of retrieved videos are shown in Table 5.
2. After the first videos retrieval, we manually selected, from that pool of videos, around 100 relevant videos that served as seeds, incorporating the YouTube API, to search for related content videos with the possible presence of hate speech content. Then, we retrieved around 390 related videos.
3. At the same time, we manually identified a few channels with publicly available playlists that may contain possible hate speech content. We downloaded some of the playlist videos and added around 2700 more videos to our pool of total videos.

---

<sup>6</sup>Our hate speech related words lists can be consulted on: <https://github.com/iltocl/dcc-hsdvmi-video-dataset.git>

<sup>7</sup><https://hatebase.org/>

<sup>8</sup><https://www.gob.mx/sre/documentos/guia-de-lenguaje-incluyente-y-no-sexista?state=published>



Video	Content	Class
	The video shows a student explicitly insulting another student because of his sexual identity	1 - Hate Speech
	The video shows two woman talking about immigrants discrimination	0 - Non-Hate Speech

Table 5: Examples of retrieved videos with possible hate speech content. Videos were retrieved from the YouTube platform after filtering them by using a spanish hate speech keywords related list.

Table 6 compile some examples of the type of videos retrieved as a result for this stage. The majority of the videos belong to stand-up shows, soap operas, news, music videos, gameplay, podcast fragments, sketches and a variety of topics. Once the videos were retrieved and downloaded, we segmented each video into 1-minute length scenes. This gave us a total of approximately 8,000 video scenes to work with.

### 5.1.2 Annotating the dataset

The video scenes were randomly grouped into batches of 250 videos. Each batch was provided to three different annotators. Where each annotator varies in age and gender from the others. Actually, we are 6 annotators, 4 men and 2 women with an age range that goes around 20’s to 40’s years. The final label was assigned in two possible ways. The first one by majority vote and the second one by considering that if at least one of the annotators labeled the video as hate speech (hate speech class corresponds to label 1) then the final label should be hate speech as well.

**Our annotation guideline.** In order to facilitate to the annotators the corresponding definitions of what is and what is not considered hate speech content we created an annotation guideline<sup>9</sup>. A general overview is shown in Figure 10. This guideline was structured in a way the annotator first decides if the video presents hate speech content (label 1) or not (label 0). If the decision is negative then the annotator should specify if the video is relevant or not. It is considered a non-relevant video as any video that is in a language different from Spanish or is a musical video without dialogues (e.g., a video with background image and instrumental music).

On the other hand, if the decision is affirmative then more questions about the video content are done. These questions consider:

<sup>9</sup>Our complete annotation guideline could be consulted on: <https://github.com/iltoocl/hsdvmi-vid-co-annotation-webapp/tree/27614f759a9a62b26baef4a6d0cdf9fe64c13fb/guideline>




Retrieved by	Videos			Type of videos
Filtered using hate-speech related words lists	 PDVJLBLEGvg	 qs2BXPib74Q	 wdgoMV1rwEg	stand-up shows soap operas news music videos
Manually selected relevant videos	 04jr6M_XS9I.03	 ajvmOU2AIWI.03	 cD8uERrn7Po.02	stand-up shows soap operas news
Related videos from relevant ones	 _aqQFPpBXO4.07	 2R-1Wiw_log.08	 CrI-9UuaFrI.08	soap operas gameplays podcast fragments variety topics
Manually identified publicly available channels and playlists	 dyvnCDvkelw.00	 MAUnbbPkb9Y.04	 cqFEnokKHGI.04	reality shows sketches stand-up shows

Table 6: Examples of retrieved videos for each step on the stage 1. *Collecting and processing videos.*

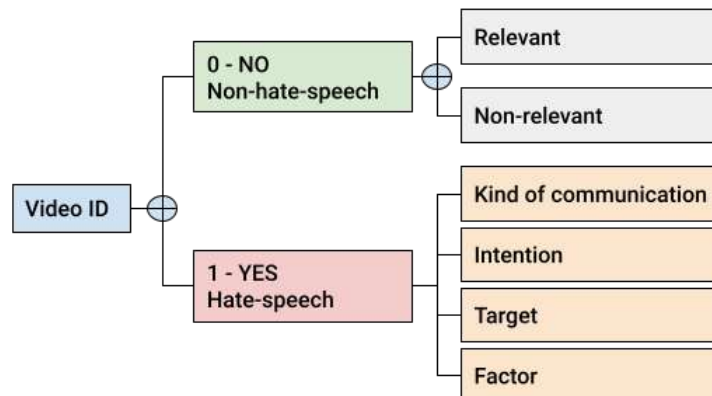


Figure 10: General overview of our annotation guidelines.

- Target. To whom the hate speech is intended, it may be an individual or a group.
- Kind of communication. The type of communication observed in the video. This involves verbal, non-verbal, visual, textual, and mass (e.g., informative news programs that report the case) types.
- Intention. The category of the intention of the message observed from the video. The

options are physical attack, disapproval or discontent expressions, and humoristic expressions.

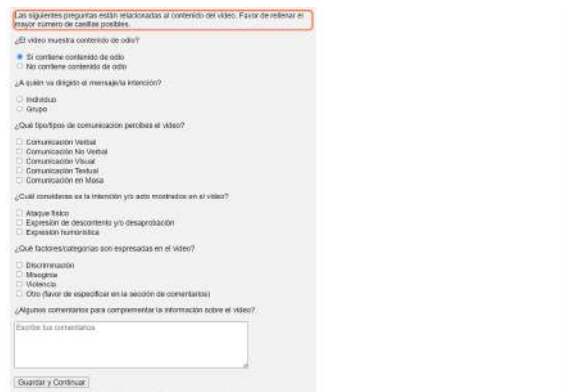
- Factor. The subtype of hate speech content. We considered discrimination, misogyny, violence, and the possibility to specify others.



(a) Home page view.



(b) Video annotation page view.



(c) Video information submit form view.

Figure 11: Overview of our annotation platform.

**Our annotation platform.** Our annotators were provided with a link to access our web app annotation platform<sup>10</sup>. They were asked to read the guidelines to understand the necessary concepts and then annotate their assigned videos. An overview of our platform is shown in Figure 11.

<sup>10</sup>Documentation of our web app annotation platform could be consulted on: <https://github.com/iltocl/hsdvmi-video-annotation-webapp.git>

### 5.1.3 Analyzing the current dataset

At this moment, we have around a little more than 250 annotated video scenes. For one batch of 250 videos, annotated by 3 annotators, an obtained average Cohen’s Kappa annotator’s agreement score of 0.45 may be interpreted as a moderate agreement, which we consider good because of the subjectivity of the hate speech detection task.

Annotator’s coincidence labeling for	Number of videos
3 videos (all assigned label 0)	169
3 videos (all assigned label 1)	20
2 videos (two assigned label 1)	14
1 videos (one assigned label 1)	47

Table 7: Number of video coincidences for annotators who worked on batch 1.

As it is shown in Table 7 the number of videos annotated as hate speech by only one annotator corresponds to the 18.8% of the total number of videos of the batch. A 5.6% corresponds to the case in were two videos coincide on the label given by two annotators and an 8% corresponds to the cases where the three annotators gave the same label (1) to a certain video. After computing the assignation of the final label by *majority vote*, we obtain 13.6% of the videos belonging to the hate speech class, this percentage increases a little if we consider *at least one vote* to assign the final label (18.8%). We consider this as a good ratio considering the annotation of 250 videos for this reference batch. Some examples of labeling coincidences for videos are shown in Table 8.




Video	Description	Assigned label by
 jifBsgwNvVQ.02	The video scene shows a woman verbally expressing discontent in a derogatory way to another woman because of her lifestyle ideology.	annotator 1 (misogyny)
 44DUP1gFp4k.02	The video scene shows two men characterized as stereotypical urban groups with a mocking intention and using derogatory language.	annotator 1 (discrimination) annotator 2 (discrimination)
 nI-czNIcqRE.03	The video scene shows a man physically attacking another man and verbally expressing derogatory adjectives related to his social status.	all annotators (discrimination, violence, discrimination)

Table 8: Examples of videos that coincide on the assignation of hate speech label (1) by the annotators. *Warning: These examples may be offensive and do not represent the perspectives of the authors.*

### 5.1.4 Formalizing our dataset

We partitioned our actual pool of annotated videos into training (80%) and test (20%). It is expected that the available annotated videos will increase with the time and help of more annotators.

According to our first version built-up dataset, two ways of evaluation will be possible for the hate speech detection task. The first one is treating it as a binary task (hate speech, non-hate speech) and the other is considered a multi-class task (discrimination, misogyny, violence, other). A summary of the actual state of our dataset is provided in Table 9.

<b>Task description</b>	Hate Speech detection (Binary: yes, no) (If YES, Type: Discrimination, Misogyny, Violence, Others)
<b>Details of the task</b>	Additional information from the “hate speech class”: Objective/Target (Individual, Group) Type of communication (Verbal, Non-verbal, Visual, Textual, Mass) Intention (Physical attack, Disapproval expression, Humoristic expression) Factor (Discrimination, Misogyny, Violence, Others)
<b>Size of the dataset</b>	8,000 videos (each video of 1-minute length)
<b>Language</b>	Spanish (Mexico)
<b>Level of annotation</b>	Video segment (each video of 1-minute length)
<b>Source</b>	YouTube
<b>Medium (modality)</b>	Video
<b>Repository</b>	Documentation process of our video dataset: <a href="https://github.com/iltocl/dcc-hsdvmi-video-dataset.git">https://github.com/iltocl/dcc-hsdvmi-video-dataset.git</a> Our annotation guideline and webapp: <a href="https://github.com/iltocl/hsdvmi-video-annotation-webapp.git">https://github.com/iltocl/hsdvmi-video-annotation-webapp.git</a> Final Spanish Mexican video dataset: <i>In process</i>

Table 9: Summary of the current state of our Mexican Spanish video dataset for the hate speech detection task.

## 5.2 Hate Speech Detection Task using Multimodal Information

Following the previously presented methodology in Section 4.2 we detail the actual advances in our multimodal method for the hate speech detection task in Mexican Spanish videos.

### 5.2.1 Extraction of multimodal features

The modalities we decided to work with are visual, audio, and text modality. Then, for each video segment the extraction of their features was done by applying a pretrained model. Reviewing state-of-the-art feature extraction models we identified four pretrained models to work with to extract the correspondent features for each modality.

Modality	Pretrained model	Extracted features
Visual	I3D [10]	flow (n*1024) RGB (n*1024)
Audio	VGGish [57]	vggish (n*128)
Text	BMT [26] Whisper [59]	captions (words & indexes) transcriptions (words & indexes)

Table 10: Pretrained models used to extract features from video content for different modalities.

Table 10 summarizes the pretrained models used for each modality and the corresponding extracted features. From visual modality, we extracted RGB and flow embedding representations. These help us to capture spatial and temporal features from the video frames. From audio modality, we extracted audio embeddings that capture patterns in the audio related to the content. In the case of text modality, we extracted both, captions and transcriptions from the video scenes. We used Whisper to transcribe the dialogues of the videos. Then, all the transcriptions are in Spanish. On the other hand, because we wanted to consider information about what was happening in the video we decided to extract captions using the BMT [26] model. This model allow us to generate an English description of each one of the video scenes. Table 11 shows examples of the generated transcriptions and captions for the text modality. It is perceived that in the majority of the cases, the generated transcriptions are highly similar to the original dialogues while the captions are not too reliable when describing the videos in most cases. This suggests us to consider captions as a temporal modality to experiment with.

Once the features are extracted we will be going to use them into the fusion stage to generate a multimodal representation that later will serve us to classify for the corresponding task.





Video	Transcription	Caption
 44DUP1gFp4k.02	<p>Estamos preparados para eso y más escabrón ¡Chairo! ¡Discurso contradictorio una vez más! ¡Chairo! Pues mira carnal, la neta Tú maldito burgués hijo de papi Tus transnacionales están jodiendo a todo el país Lo de hoy tiene que ser el truke, la verdad, la neta, la neta ¿Qué? Ah, disculpa príncipe Es que estaba hablando con mi rey que ya se armó el viaje a Capulquiri Si quieres ir con toda confianza, eh Ya tenemos el yate, ya está todo arreglado Es más, ¿qué tan bueno eres para servir tragos? Mira, te volamos la mata, te ponemos ropita y ¡huy! Mi papi te puede pagar muy bien si nos atiendes padre, ¿qué dices? Ah... ¡Chairo, usa descalificación! P*** puto Ja, mi rey, usa guarura ¡Opeye! Quitame este gato No, carnal, espérate ¿Qué? ¿Qué te pasa, carnal? ¡No! ¡No! ¡No! ¡Pero qué cobre esto, choro! ¡No! ¡No! ¡No está en 43, carnal! ¡Vamos por 44! ¡No seas culero, mi Pokémon! ¡Uy, mi rey! Espero hay un centro Pokémon cerca porque se lo va a llevar la verga</p>	<p>a man is seen speaking to the camera and leads into him walking around a park a man is seen speaking to the camera while holding a stick a man is seen walking around a street and holding a stick and walking around a house a woman is standing in a street talking he then walks away and walks away and walks away</p>
 nL-czNlcqRE.03	<p>la tenía engañada eres tú. Pero sabes que ya me hartaste. Le voy a decir a mi suegro que te ponga de patitas en la calle. Naco igualado. Yo no me voy a rebajar a tu nivel. Qué bueno que la diseñadora pudo tener el vestido para esta fecha. Sí, mamá. Te dejamos descansar. Mañana será el gran día. ¿Tú todavía tienes esta noche para pensar con el corazón? Si no estás segura, no te cases. Pero ¿qué consejos son estos, mamá? Dejemos a Julieta sola y no la confundamos. Buenas noches, mi vida. ¿Cómo te atreves a dar un consejo así una noche antes de la boda, mamá? ¿Pero qué te pasa?</p>	<p>two men are seen sitting on a table and leads into them playing a game of beer pong the two continue talking and the man in a room and leads into a woman speaking to the camera the man and the man are seen speaking to the camera and leads into them walking away a man in a black shirt is standing in a room the woman then grabs a towel and puts a towel on the woman and puts her hands in the end a man in a black shirt is standing in front of a microphone</p>

Table 11: Example of captions and transcription obtained for videos annotated as hate speech content. *Warning: These examples may be offensive and do not represent the perspectives of the authors.*

### 5.3 Preliminary Conclusions

According to our proposed methodology in Section 4 our preliminary results are divided into advances related to our proposed Mexican Spanish video dataset for the hate speech detection task (Section 4.1, Section 5.1), and advances related to the multimodal proposed method (Section 4.2, Section 5.2).

In section 5.1 we detailed the steps we applied to obtain the first annotated subset version of our video dataset. This process included the collection and processing of videos with possible related hate-speech content; the definition, and creation of a guideline and annotation platform to annotate the previously retrieved videos; a description and analysis of the current subset of annotated videos and the details of how we plan to formalize our dataset. Because we already covered the general pipeline of dataset creation, we expect that in the next months, our annotated dataset will increase by inviting more annotators to participate. Thus, we will actualize the statistics of our dataset and refine it with each annotated batch of videos.

On the other hand, in section 5.2, related to the multimodal proposed method, we described the way we extracted visual (flow and RGB vectors), audio (vggish vectors), and text (word embeddings from captions and transcriptions) features directly from our subset of annotated videos. According to our methodology (Section 4.2) the next step will be the

experimentation with the modality representations by exploring our proposed ideas for the fusion method (Section 4.2.2) in order to generate a valuable multimodal representation of the content. The ideas include experimenting with a variation of Locally Linear Embedding (LLE) oriented to multimodal embeddings, learning the multimodal embeddings, or using multiple kernels learning to fusion the modalities. It is planned that this representation serves as input to an attention-based model to capture relevant information that enhances the performance in the classification stage. Then, vanilla transformer architecture and its variations will be considered for the experimentation. Particularly, the task is intended to be evaluated as a binary task (the video scene contains hate speech or does not contain hate speech) but there is also a possibility to expand the evaluation to a multi-class approach because we already considered some categories of hate speech (discrimination, misogyny, violence, other) when we established the annotation guideline.

## 6 Final Remarks

Because of the exponential growth in the use of social media and content sharing and generation, competent ways of monitoring the content are required. Despite the remarkable progress of existing state-of-the-art approaches for the hate speech detection task, we identified two research opportunities. First, existing multimodal models for the task are aimed to work mainly on text and visual modalities. This implies there is a lack of adaptation of models that analyze complex content like videos by considering the use of other modalities like audio. Second, there is still a lack of research on non-English content.

Therefore, there are two significant contributions that this work aims to obtain. First, we expect to provide a novel multimodal method that achieves competitive performance for the hate speech detection task in video content by leveraging the use of the available modalities information. Second, we expect to address the gap in research for Spanish language content for the task by constructing a Mexican Spanish language-based video dataset that will serve as a reference for future studies in the field.

At the moment, we already presented the first version of our Mexican Spanish video dataset (Section 5.1) related to the hate speech detection task and a few advances on the stages towards our multimodal method proposal (Section 5.2). Future general activities include the expansion and refinement of our video dataset, as well as, experimentation and adaptation of our method. The details are described in the following Section 6.1.

### 6.1 Future Work

To sum up, some of the main concerns we identified in the use of multimodal information, considering videos as a source, are related to the extraction, fusion, and use of the multiple available modalities by representing their relation and interactions into a shared space. Moreover, additional factors like the consideration and inclusion of temporal information into the shared representation also represent a research opportunity.

While most works are oriented to work with one or two modalities for video content, we aim to leverage the representation of  $n$  different modalities, including audio, visual, and text (captioning, transcription) into a shared representation. It is expected that this multimodal representation successfully captures the relevant information and interaction between modalities. To achieve it three main ideas are being considered and will be explored: extending a Locally Linear Embedding (LLE) for multimodal representations, learning of the multimodal representations, and using an external kernel to fusion modalities. Thus, by using the multimodal representation as input to an attention-based model we aim to leverage the distinction between videos with possible hate speech content and videos without hate speech content.

Additionally, while most works are oriented to English resources, we highlight our advances in our Mexican Spanish video dataset oriented to the hate speech detection task. The main future work activities include extending the dataset by inviting more annota-

tors to the project, refining it with the new annotated batch statistics, formalizing it by documenting the process, and releasing and experimenting on a large version.

## 6.2 Work Plan

To follow up, Figure 12 summarizes the estimated timelines for future work activities.

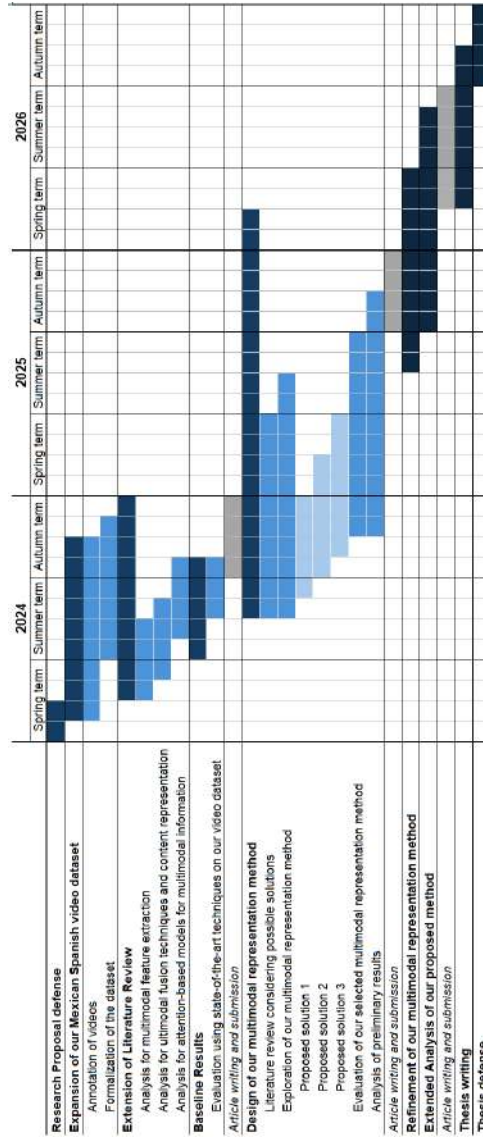


Figure 12: Future work plan activities overview.

## Bibliography

- [1] Hassan Akbari et al. “VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text”. In: *Neural Information Processing Systems*. 2021.
- [2] Cleber Alcântara, Viviane Pereira Moreira, and Diego de Vargas Feijó. “Offensive Video Detection: Dataset and Baseline Results”. In: *International Conference on Language Resources and Evaluation*. 2020.
- [3] X. Amatriain. “Transformer models: an introduction and catalog”. In: *ArXiv abs/2302.07730* (2023).
- [4] Vishal Anand et al. “Customized video filtering on YouTube”. In: *ArXiv abs/1911.04013* (2019).
- [5] Anurag Arnab et al. “ViViT: A Video Vision Transformer”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 6816–6826.
- [6] Simon Baker et al. “A Database and Evaluation Methodology for Optical Flow”. In: *2007 IEEE 11th International Conference on Computer Vision*. 2007, pp. 1–8. DOI: 10.1109/ICCV.2007.4408903.
- [7] Tadas Baltruaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal Machine Learning: A Survey and Taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41 (2017), pp. 423–443. URL: <https://api.semanticscholar.org/CorpusID:10137425>.
- [8] Valerio Basile et al. “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter”. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: <https://aclanthology.org/S19-2007>.
- [9] Yoshua Bengio et al. “A Neural Probabilistic Language Model”. In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435.
- [10] João Carreira and Andrew Zisserman. “Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 4724–4733. URL: <https://api.semanticscholar.org/CorpusID:206596127>.
- [11] Anusha Chhabra and Dinesh Kumar Vishwakarma. “A literature survey on multimodal and multilingual automatic hate speech identification”. In: *Multimedia Systems* (2023).
- [12] Anusha Chhabra and Dinesh Kumar Vishwakarma. “A literature survey on multimodal and multilingual automatic hate speech identification”. In: *Multimedia Systems* 29.3 (Jan. 2023), pp. 1203–1230. ISSN: 1432-1882. DOI: 10.1007/s00530-023-01051-8. URL: <http://dx.doi.org/10.1007/s00530-023-01051-8>.

- [13] Alexey Dosovitskiy et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. In: *ArXiv abs/2010.11929* (2020). URL: <https://api.semanticscholar.org/CorpusID:225039882>.
- [14] Christoph Feichtenhofer et al. “SlowFast Networks for Video Recognition”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2018), pp. 6201–6210. URL: <https://api.semanticscholar.org/CorpusID:54463801>.
- [15] Elisabetta Fersini et al. “SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification”. In: *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 533–549. DOI: 10.18653/v1/2022.semeval-1.74. URL: <https://aclanthology.org/2022.semeval-1.74>.
- [16] Paula Fortuna and Sérgio Nunes. “A Survey on Automatic Detection of Hate Speech in Text”. In: *ACM Computing Surveys (CSUR)* 51 (2018), pp. 1–30.
- [17] Zhe Gan et al. 2022.
- [18] Jing Gao et al. “A Survey on Deep Learning for Multimodal Data Fusion”. In: *Neural Computation* 32.5 (May 2020), pp. 829–864. ISSN: 1530-888X. DOI: 10.1162/neco\_a\_01273. URL: [http://dx.doi.org/10.1162/neco\\_a\\_01273](http://dx.doi.org/10.1162/neco_a_01273).
- [19] Jort F. Gemmeke et al. “Audio Set: An ontology and human-labeled dataset for audio events”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.
- [20] *Global Social Media Statistics - DataReportal – global digital insights*. URL: <https://datareportal.com/social-media-users>.
- [21] Raul Gomez et al. “Exploring Hate Speech Detection in Multimodal Publications”. In: Mar. 2020, pp. 1459–1467. DOI: 10.1109/WACV45572.2020.9093414.
- [22] Mehmet Gönen and Ethem Alpaydin. “Multiple Kernel Learning Algorithms”. In: *Journal of Machine Learning Research* 12.64 (2011), pp. 2211–2268. URL: <http://jmlr.org/papers/v12/gonen11a.html>.
- [23] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.
- [24] Shawn Hershey et al. “CNN Architectures for Large-Scale Audio Classification”. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. URL: <https://arxiv.org/abs/1609.09430>.
- [25] Tak-Wai Hui, Xiaou Tang, and Chen Change Loy. “LiteFlowNet: A Lightweight Convolutional Neural Network for Optical Flow Estimation”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 8981–8989. URL: <https://api.semanticscholar.org/CorpusID:29162783>.

- [26] Vladimir E. Iashin and Esa Rahtu. “A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer”. In: *ArXiv* abs/2005.08271 (2020). URL: <https://api.semanticscholar.org/CorpusID:218674428>.
- [27] Eddy Ilg et al. “FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 1647–1655. URL: <https://api.semanticscholar.org/CorpusID:3759573>.
- [28] Rommel Hernandez Urbano Jr. et al. “A BERT-based Hate Speech Classifier from Transcribed Online Short-Form Videos”. In: *2021 5th International Conference on E-Society, E-Education and E-Technology* (2021).
- [29] Uday Kamath, John Liu, and James Whitaker. *Deep Learning for NLP and Speech Recognition*. 1st. Springer Publishing Company, Incorporated, 2019. ISBN: 3030145956.
- [30] Hitesh Kumar Sharma, K Kshitiz, and Shailendra. “NLP and Machine Learning Techniques for Detecting Insulting Comments on Social Networking Platforms”. In: *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. 2018, pp. 265–272. DOI: 10.1109/ICACCE.2018.8441728.
- [31] Ashok Kumar L, Karthika Renuka D, and Shunmuga Priya M.C. “Analysis of Audio Visual Feature Extraction Techniques for AVSR System”. In: EAI, Dec. 2021. DOI: 10.4108/eai.7-12-2021.2314528.
- [32] Sangho Lee et al. “Parameter Efficient Multimodal Transformers for Video Representation Learning”. In: *International Conference on Learning Representations (ICLR)*. May 2021. arXiv: 2012.04124.
- [33] Zheng Lian et al. “Investigation of Multimodal Features, Classifiers and Fusion Methods for Emotion Recognition”. In: *ArXiv* abs/1809.06225 (2018). URL: <https://api.semanticscholar.org/CorpusID:52284616>.
- [34] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. “Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions”. In: 2022. URL: <https://api.semanticscholar.org/CorpusID:257038528>.
- [35] Morency Louis-Philippe and Llang Paul. *Tutorial on MultiModal Machine Learning*. Online tutorial. 2023. URL: <https://cmu-multicomp-lab.github.io/mml-tutorial/icml2023/>.
- [36] Sijie Mai et al. “Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis”. In: *IEEE Transactions on Affective Computing* (2022), pp. 1–1. DOI: 10.1109/TAFFC.2022.3172360.
- [37] Mishaim Malik et al. “Automatic speech recognition: a survey”. In: *Multimedia Tools and Applications* 80.6 (Nov. 2020), pp. 9411–9457. ISSN: 1573-7721. DOI: 10.1007/s11042-020-10073-7. URL: <http://dx.doi.org/10.1007/s11042-020-10073-7>.

- [38] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
- [39] Oumaima Moutik et al. “Convolutional Neural Networks or Vision Transformers: Who Will Win the Race for Action Recognitions in Visual Data?” In: *Sensors 23* (Jan. 2023), p. 734. DOI: 10.3390/s23020734.
- [40] David Olson and Dursun Delen. *Advanced Data Mining Techniques*. Springer, Jan. 2008. ISBN: 978-3-540-76916-3.
- [41] M. Patrick et al. “Space-Time Crop amp; Attend: Improving Cross-modal Video Representation Learning”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2021, pp. 10540–10552. DOI: 10.1109/ICCV48922.2021.01039. URL: <https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01039>.
- [42] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: <https://aclanthology.org/D14-1162>.
- [43] Fabio Poletto et al. “Resources and benchmark corpora for hate speech detection: a systematic review”. In: *Language Resources and Evaluation 55* (2020), pp. 477–523.
- [44] Aneri Rana and Sonali Jha. “Emotion Based Hate Speech Detection using Multi-modal Learning”. In: *ArXiv abs/2202.06218* (2022).
- [45] Sam T. Roweis and Lawrence K. Saul. “Nonlinear Dimensionality Reduction by Locally Linear Embedding”. In: *Science 290.5500* (2000), pp. 2323–2326. DOI: 10.1126/science.290.5500.2323. eprint: <https://www.science.org/doi/pdf/10.1126/science.290.5500.2323>. URL: <https://www.science.org/doi/abs/10.1126/science.290.5500.2323>.
- [46] Anna Schmidt and Michael Wiegand. “A Survey on Hate Speech Detection using Natural Language Processing”. In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. DOI: 10.18653/v1/W17-1101. URL: <https://aclanthology.org/W17-1101>.
- [47] J. Selva et al. “Video Transformers: A Survey”. In: *IEEE Transactions on Pattern Analysis amp; Machine Intelligence 45.11* (Nov. 2023), pp. 12922–12943. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2023.3243465.
- [48] Lanyu Shang et al. “VulnerCheck: A Content-Agnostic Detector for Online Hatred-Vulnerable Videos”. In: Dec. 2019, pp. 573–582. DOI: 10.1109/BigData47090.2019.9006329.



- [49] Chhavi Sharma et al. “SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!” In: *ArXiv abs/2008.03781* (2020).
- [50] Bowen Shi et al. “Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction”. In: *ArXiv abs/2201.02184* (2022). URL: <https://api.semanticscholar.org/CorpusID:245769552>.
- [51] Urmila Shrawankar and Vilas M. Thakare. “Techniques for Feature Extraction In Speech Recognition System : A Comparative Study”. In: *ArXiv abs/1305.1145* (2013). URL: <https://api.semanticscholar.org/CorpusID:5473330>.
- [52] Chen Sun et al. “Contrastive Bidirectional Transformer for Temporal Representation Learning”. In: *ArXiv abs/1906.05743* (2019).
- [53] M. Suresha, Subramanya Kuppa, and D. S. Raghukumar. “A study on deep learning spatiotemporal models and feature extraction techniques for video understanding”. In: *International Journal of Multimedia Information Retrieval* 9 (2020), pp. 81–101. URL: <https://api.semanticscholar.org/CorpusID:210865135>.
- [54] *The most spoken languages in the world*. URL: <https://www.berlitz.com/blog/most-spoken-languages-world>.
- [55] Du Tran et al. “A Closer Look at Spatiotemporal Convolutions for Action Recognition”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 6450–6459. URL: <https://api.semanticscholar.org/CorpusID:206596999>.
- [56] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [57] *VGGish*. URL: <https://github.com/tensorflow/models/blob/0b3a8abf095cb8866ca74c2e118c1894c0e6f947/research/audioset/vggish/README.md>.
- [58] *What is hate speech?* URL: <https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech>.
- [59] *Whisper*. URL: <https://github.com/openai/whisper.git>.
- [60] Ching Seh Wu and Unnathi Bhandary. “Detection of Hate Speech in Videos Using Machine Learning”. In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. 2020, pp. 585–590. DOI: 10.1109/CSCI51800.2020.00104.
- [61] Saining Xie et al. “Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification”. In: *European Conference on Computer Vision*. 2017. URL: <https://api.semanticscholar.org/CorpusID:51863579>.

- [62] Fan Yang et al. “Exploring Deep Multimodal Fusion of Text and Photo for Hate Speech Classification”. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 11–18. DOI: 10.18653/v1/W19-3502. URL: <https://aclanthology.org/W19-3502>.
- [63] Marcos Zampieri et al. “SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1425–1447. DOI: 10.18653/v1/2020. semeval-1.188. URL: <https://aclanthology.org/2020.semeval-1.188>.
- [64] Su-Fang Zhang et al. “Multimodal Representation Learning: Advances, Trends and Challenges”. In: *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*. 2019, pp. 1–6. DOI: 10.1109/ICMLC48188.2019.8949228.
- [65] Alex Zihao Zhu et al. “EV-FlowNet: Self-Supervised Optical Flow Estimation for Event-based Cameras”. In: *ArXiv* abs/1802.06898 (2018). URL: <https://api.semanticscholar.org/CorpusID:3396150>.
- [66] Qiu-shi Zhu et al. “VATLM: Visual-Audio-Text Pre-Training with Unified Masked Prediction for Speech Representation Learning”. In: *ArXiv* abs/2211.11275 (2022).