MULTIBENCH & MULTIZOO
Resource Description
Tutorial

Presented by:
Itzel Tlelo
Arnold Morales

June, 2024

Instituto Nacional de Astrofísica, Óptica y Electrónica
Computer Science Department

Introduction

Our perception of the natural world surroundings us **involves multiple sensory modalities**: we see objects, hear audio signals, feel textures, smell fragrances, and taste flavors.

A **modality** refers to **a way in which a signal exists or is experienced**. Multiple modalities then refer to a combination of multiple signals each expressed in heterogeneous manners.

Learning multimodal representations involve **integrating information** from **multiple** heterogeneous **sources** of data.

It may be considered a **challenging yet crucial area** with **numerous real-world applications** in multimedia, affective computing, robotics, finance, human-computer interaction, and healthcare.

Limitations of current multimodal datasets

Typically focus on performance without quantifying the potential drawbacks involved with:

- **time**
- **space complexity**
- **robustness**

In real-world applications a **balance** between **performance, robustness, and complexity is often required**

MULTI BENCH

It was released in order to:

- **accelerate progress towards understudied modalities and tasks** while ensuring real-world robustness

Milestone in unifying disjoint efforts in multimodal machine learning research

Paves the way towards a better understanding of the capabilities and limitations of multimodal models, all the while **ensuring:**

- **ease of use**
- **accessibility**
- **reproducibility**

MULTI BENCH

A **systematic** and **unified large-scale benchmark** for **multimodal learning**.

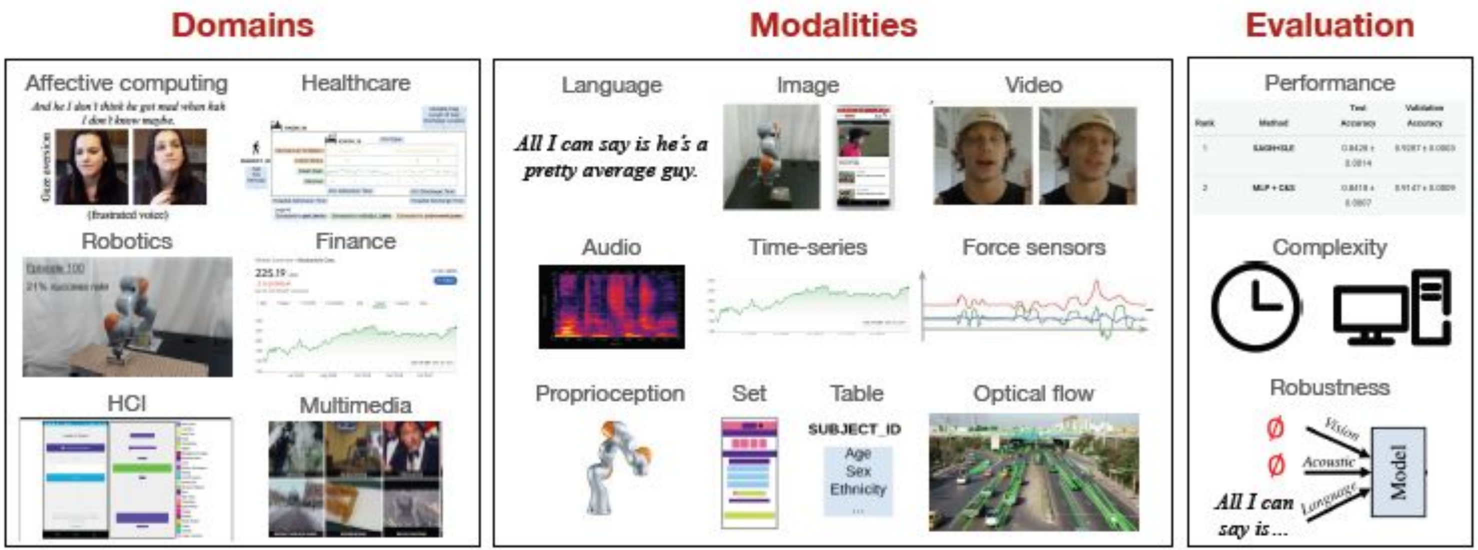


Figure 1: MULTIBENCH contains a diverse set of 15 datasets spanning 10 modalities and testing for more than 20 prediction tasks across 6 distinct research areas, and enables standardized, reliable, and reproducible large-scale benchmarking of multimodal models for performance, complexity, and robustness.

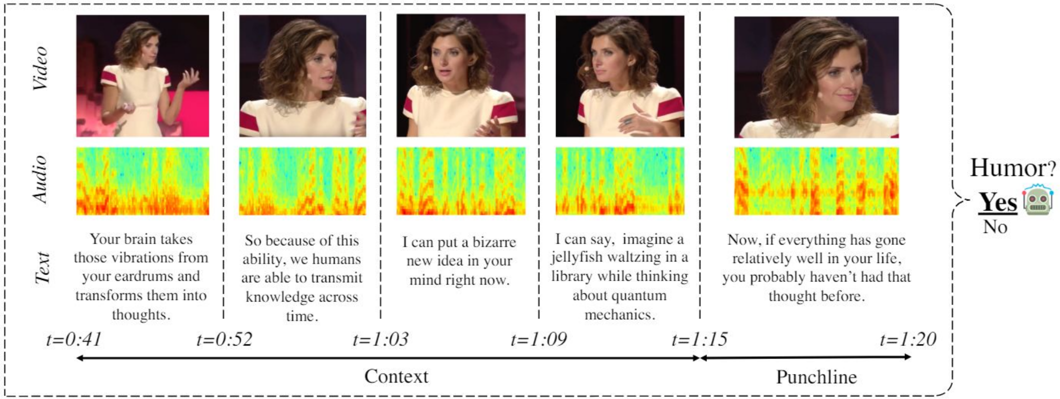
Datasets

Table 1: MULTIBENCH provides a comprehensive suite of 15 multimodal datasets to benchmark current and proposed approaches in multimodal representation learning. It covers a diverse range of research areas, dataset sizes, input modalities (in the form of ℓ: language, i: image, v: video, a: audio, t: time-series, ta: tabular, f: force sensor, p: proprioception sensor, s: set, o: optical flow), and prediction tasks. We provide a standardized data loader for datasets in MULTIBENCH, along with a set of state-of-the-art multimodal models.

Research Area	Size	Dataset	Modalities	# Samples	Prediction task
Affective Computing	S	MUSTARD [24]	$\{\ell, v, a\}$	690	sarcasm
	M	CMU-MOSI [181]	$\{\ell, v, a\}$	2,199	sentiment
	L	UR-FUNNY [64]	$\{\ell, v, a\}$	16,514	humor
	L	CMU-MOSEI [183]	$\{\ell, v, a\}$	22,777	sentiment, emotions
Healthcare	L	MIMIC [78]	$\{t, ta\}$	36,212	mortality, ICD-9 codes
Robotics	M	MUJOCO PUSH [90]	$\{i, f, p\}$	37,990	object pose
	L	VISION&TOUCH [92]	$\{i, f, p\}$	147,000	contact, robot pose
Finance	M	STOCKS-F&B	$\{t \times 18\}$	5,218	stock price, volatility
	M	STOCKS-HEALTH	$\{t \times 63\}$	5,218	stock price, volatility
	M	STOCKS-TECH	$\{t \times 100\}$	5,218	stock price, volatility
HCI	S	ENRICO [93]	$\{i, s\}$	1,460	design interface
Multimedia	S	KINETICS400-S [80]	$\{v, a, o\}$	2,624	human action
	M	MM-IMDB [8]	$\{\ell, i\}$	25,959	movie genre
	M	AV-MNIST [161]	$\{i, a\}$	70,000	digit
	L	KINETICS400-L [80]	$\{v, a, o\}$	306,245	human action

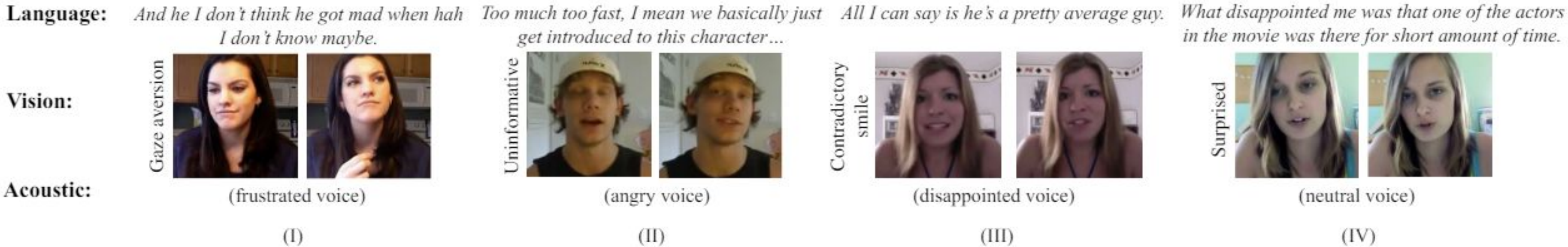
MUSTARD

UR-FUNNY



CMU-MOSI

CMU-MOSEI



Dataset		Language	Vision	Audio	Prediction task
MUSTARD	YouTube TV shows	Text utterances BERT representations GloVe word vectors	Visual features (frames) pool5 layer of ImageNet pretrained ResNet-152 model Facial expression features OpenFace	Low-level features Librosa library COVAREP software	**sarcasm** sarcastic non-sarcastic
CMU-MOSI	YouTube Opinion	Transcripts GloVe word embeddings	Visual features (full video segment) Facet library (facial action units, facial landmarks, head pose, gaze tracking, HOG features) Facial expression features Open Face	Acoustic features COVAREP software (12 mel-frequency, pitch tracking, voiced/unvoiced segment features)	**sentiment** sentiment intensity [-3,+3]
UR-FUNNY	TED talks Humorous punchlines	Transcripts GloVe word embeddings	same as CMU-MOSI	same as CMU-MOSI	**humor** binary
CMU-MOSEI	YouTube Opinion	same as CMU-MOSI	same as CMU-MOSI	same as CMU-MOSI	**sentiment, emotions** 9 discrete emotions (angry, excited, fear, sad, surprised, frustrated, happy, disappointed, and neutral) continuous emotions (valence, arousal, and dominance)

MULTI ZOO (MULTI BENCH toolkit)

Figure 2: Our MULTIBENCH toolkit provides a machine learning pipeline across data processing, data loading, multimodal models, evaluation metrics, and a public leaderboard to encourage accessible, standardized, and reproducible research in multimodal representation learning.

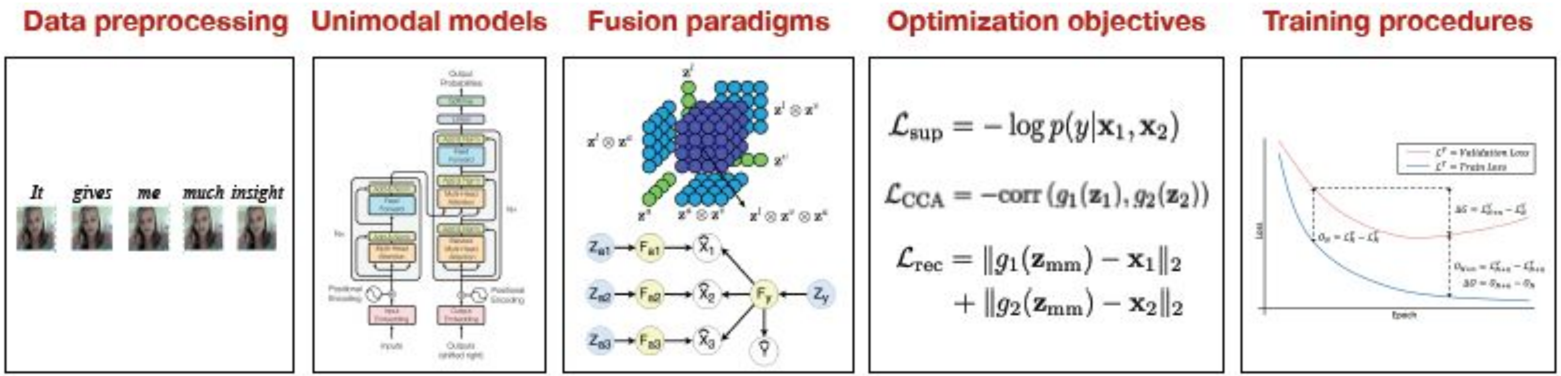


Figure 3: MULTIZOO provides a standardized implementation of multimodal methods in a modular fashion to enable accessibility for new researchers, compositionality of approaches, and reproducibility of results.

MULTI ZOO (MULTI BENCH toolkit)

Table 2: MULTIZOO provides a standardized implementation of the following multimodal methods spanning data processing, fusion paradigms, optimization objectives, and training procedures, which offer complementary perspectives towards tackling multimodal challenges in alignment, complementarity, and robustness.

Category	Method	Alignment	Complementarity	Robustness
Data	WORDALIGN (Chen et al., 2017)	✓	✗	✗
Model	EF, LF (Baltrušaitis et al., 2018)	✗	✓	✗
	TF (Zadeh et al., 2017), LRTF (Liu et al., 2018)	✗	✓	✗
	MI-MATRIX, MI-VECTOR, MI-SCALAR (Jayakumar et al., 2020)	✗	✓	✗
	NL GATE (Wang et al., 2020)	✗	✓	✗
	MULT (Tsai et al., 2019a)	✓	✓	✗
	MFAS (Pérez-Rúa et al., 2019)	✗	✓	✗
Objective	CCA (Andrew et al., 2013)	✓	✗	✗
	REFNET (Sankaran et al., 2021)	✓	✗	✗
	MFM (Tsai et al., 2019b)	✗	✓	✗
	MVAE (Wu and Goodman, 2018)	✗	✓	✗
	MCTN (Pham et al., 2019)	✗	✗	✓
Training	GRADBLEND (Wang et al., 2020)	✗	✓	✓
	RMFE (Gat et al., 2020)	✗	✓	✓

Evaluation Protocol

- **Performance (standardized evaluation metrics designed for each dataset)**:
  - MSE and MAE for regression
  - accuracy, micro & macro F1- score, and AUPRC for classification
- **Complexity**:
  - amount of information taken in bits (i.e., data size)
  - number of model parameters
  - time and memory resources (during the entire training process)
  - inference time
  - memory on CPU and GPU
- **Robustness**
  - **Modality-specific imperfections** applied to each modality taking into account its unique noise topologies
  - **Multimodal imperfections** capture correlations in imperfections across modalities (e.g., missing modalities, or a chunk of time missing in multimodal time-series data)

Final Remarks

MULTI BENCH

A large-scale **multimodal benchmark**

- **Focus on ease of use, accessibility, and reproducibility**
- Involves a much more **diverse set of modalities** (e.g., tabular data, time-series, sensors, graph and set data) **and tasks**
- **Evaluates performance, complexity and robustness**
- Searches for the **standardization of multimodal learning**

MULTI ZOO

A **multimodal toolkit**

- For **building more generalizable, lightweight, and robust multimodal models**
- **Publicly available**
- **Regularly updated** with new tasks and modeling paradigms
- Welcome inputs from the community

Final Remarks

Limitations

- **Tradeoffs** between **generality and specificity**
  - desirable to build models that work across modalities and tasks
  - merit in building modality and task-specific models
- **Scale** of datasets, models, and metrics

Projected expansions

- Other multimodal research problems
- New evaluation metrics
- Multimodal transfer learning and co-learning
- Multitask learning across modalities

References

Liang, P.P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M.A., Zhu, Y., Salakhutdinov, R., & Morency, L. (2021). MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. Advances in neural information processing systems, 2021 DB1, 1-20 .

Liang, P.P., Lyu, Y., Fan, X., Agarwal, A., Cheng, Y., Morency, L., & Salakhutdinov, R. (2023). MultiZoo & MultiBench: A Standardized Toolkit for Multimodal Deep Learning. ArXiv, abs/2306.16413.

Multibench and Multizoo Source Code available at: https://github.com/pliang279/MultiBench

Zadeh, A., Zellers, R., Pincus, E., & Morency, L. (2016). MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. ArXiv, abs/1606.06259.

Zadeh, A., Liang, P.P., Poria, S., Cambria, E., & Morency, L. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Annual Meeting of the Association for Computational Linguistics.

Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). Annual Meeting of the Association for Computational Linguistics.

Hasan, M., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M., Morency, L., & Hoque, E. (2019). UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. Conference on Empirical Methods in Natural Language Processing.

Tutorial Code Samples

Source Code available at:
https://github.com/pliang279/MultiBench

Source Documentation available at:
https://multibench.readthedocs.io/en/latest/

Our Tutorial Code available at:
https://github.com/iltocl/dcc-tutorial-multizoo-multibench

Comic Mischief

(a) Gory Humor (b) Slapstick

(c) Mature Humor (d) Sarcasm

Figure 1: Examples of comic mischief content in movi

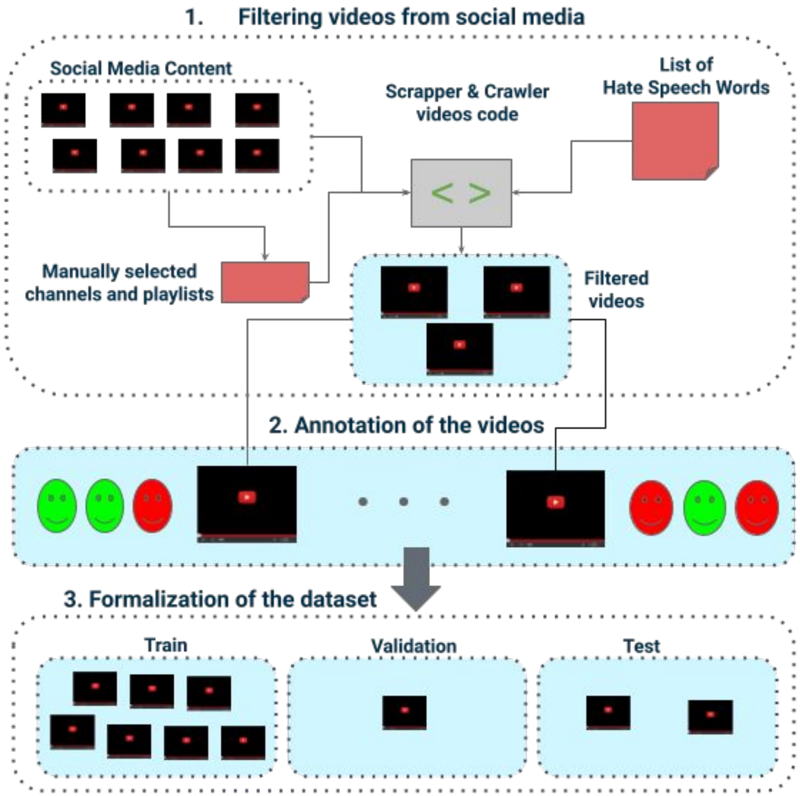
HateSpeech

Video	Content	Class
 jifBsgwNvVQ.02	The video scene shows a woman verbally expressing discontent in a despective way to another woman because of her lifestyle ideology	1 - Hate Speech (misogyny)
 44DUP1gFp4k.02	The video scene shows two men characterized as stereotypical urban groups with a mocking intention and using respective language	1 - Hate Speech (discrimination)
 nlczNlcqRE.03	The video scene shows a man physically attacking another man and verbally expressing despective adjectives related to the other man social status	1 - Hate Speech (violence)
 fVz_LEKUWrw.00	The video scene shows an informative video about psychology related concepts	0 - Non-Hate Speech

Table 3. Screenshots examples of labeled videos. Warning: These samples may be offensive and do not represent the perspectives of the authors.

Dataset		Language	Vision	Audio	Prediction task
Comic Mischief	YouTube TV shows	Captions BERT	Visual features (frames) I3D (flow, RGB)	Low-level features VGGish	**comic mischief** Binary Multilabel (gory, slapstick, mature, sarcasm)
HateSpeech	YouTube Opinion TV shows	Transcripts BERT	Visual features (full video segment) I3D (flow, RGB)	Acoustic features VGGish	**hate speech** Binary

Towards a Dataset for Hate Speech Detection in Videos



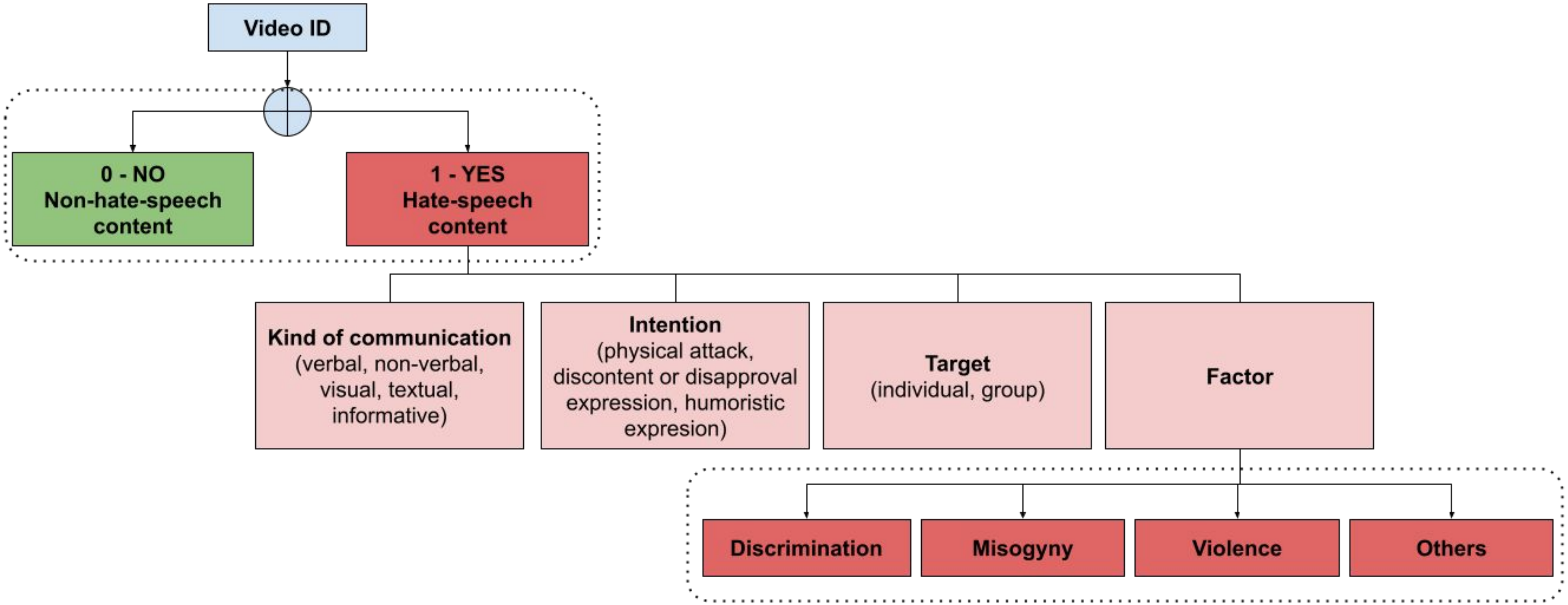
1. Filtering videos from social media

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_1og.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

Each video was segmented into **1-minute length scenes**.
This gave us a total of approximately **8,000 video scenes**.

*https://github.com/iltocl/dcc-hsdvmi-video-dataset/tree/8849f1a60e207f78f5a100937b1b27a5438a44d5/1.1%20hate%20speech%20word%20lists

2. Annotation of the videos



*https://github.com/iltocl/hsdvmi-video-annotation-webapp/blob/55a47b41f166279ae03b1263daf822090a40aeed/guideline/gu%C3%ADa%20de%20anotaci%C3%B3n%20para%20dataset%20de%20videos%20de%20discurso%20de%20odio.md

Examples of annotated videos

Video	Description	Assigned label by
 jifBsgwNvVQ.02	The video scene shows a woman verbally expressing discontent in a despective 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 despective language.	annotator 1 (discrimination) annotator 2 (discrimination)
 nl-czNIcqRE.03	The video scene shows a man physically attacking another man and verbally expressing despective adjectives related to the another man social status.	all annotators (discrimination, violence, discrimination)

First annotated subset of videos

Class	Train	Validation	Test
Hate Speech	56	9	16
No Hate Speech	118	18	33
Total	174	27	49