Privacy VITA: a new multilingual and multimodal annotated video corpus to evaluate anonymization systems

Jorge Rico*, Sofia Contreras, Enrique Manjavacas Arevalo, Maria Viana
Ruben Pérez-Ramón, Maria Luisa Izquierdo, María José Vilella, Jaime Corton
Silvia Rodriguez, Fernando Espinza, Rafael Ginard, César Pérez
Jesús Arias, Richard Cook, Pablo Regodón, Manuel Moyano
Ricardo Heredia, Lara De Santos, Pierre Plaza, Jacqueline González

Sigma Cognition
Av. de Manoteras 6, 28050 Madrid, Spain

Abstract

In this paper, we introduce a new multimodal and multilingual public resource for evaluating anonymization models and solutions, comprising a collection of 516 annotated videos. Unlike previously available resources, the Privacy VITA corpus is both multilingual and multimodal, covering Video, Image, Text and Audio modalities. The development of this dataset is motivated by the growing need to anonymize private data in an increasingly multimedia-driven society. Furthermore, emerging regulations pose significant challenges for organizations and companies that manage sensitive information while ensuring legal compliance. This paper details the processes of collection, preprocessing, annotation, and curation of the corpus, as well as its overall scope. We hope this resource will contribute to the advancement of multimodal anonymization research.

1 Introduction

The preservation and storage of personal data have become an increasingly important matter for both individuals and organizations striving to comply with evolving legal frameworks (Warren and Brandeis, 1890). Early concerns about the appropriate handling of personal information can be found in the 1974 US Privacy Act (Electronic Privacy Information Center (EPIC), n.d.), which defined the conditions under which data could be shared with third parties. Building on this foundation, successive regulations have progressively strengthened data protection standards, culminating in the European Union's General Data Protection Regulation (GDPR, European Parliament and Council, 2016), adopted in 2016 and enforced in 2018.

The need for stronger privacy management has been underscored by numerous large-scale data breaches, with an average global cost of $4.24 million per incident in 2021 (IBM Security, 2021). Incidents such as the illicit harvesting of 87 million Facebook profiles by Cambridge Analytica (Cadwalladr and Graham-Harrison, 2018) or the Yahoo! breach affecting 500 million accounts highlight the urgency of developing robust mechanisms to safeguard personal information (Verizon Business, 2023). Yet, merely storing, encrypting, or restricting access to data is insufficient. Organizations frequently need to share datasets for legitimate purposes while maintaining privacy, making anonymization a key strategy to preserve data utility while mitigating risk (Article 29 Data Protection Working Party, 2014; El Emam and Dankar, 2008).

The corpus introduced in this paper is intended to serve as an evaluation resource for systems pursuing this objective.

2 Related Work

The study of data anonymization has enabled a wide variety of corpora and resources. Although the first efforts focused on tabular and text data, recent years have seen growing interest in extending these methods to audiovisual media.

Anonymization of structured data traces its origins to the *Adult* dataset (Becker and Kohavi, 1996) which laid the foundations for classic privacy metrics such as k-anonymity, l-diversity and t-closeness. In the textual domain, notable examples include the MAPA (Arranz et al., 2022) and the TAB Benchmark (Pilán et al., 2022), aimed at detecting and masking of sensitive entities using named entity recognition (NER) techniques. MAPA provides a multilingual corpus with detailed annotations in legal and medical contexts, whereas TAB focuses on judicial texts from the European Court of Human Rights.

In the audiovisual domain, research has primarily focused on the protection of visual identity. Datasets such as AViD (Piergiovanni and Ryoo, 2020) and PA-HMDB51 (Wu et al., 2022) include

*Corresponding author: jrico@sigmacognition.ai.

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

videos of human actions with anonymized faces or privacy annotations at the frame level, while FIVA (Rosberg et al., 2023) adopts a more technical approach based on controlled facial manipulation to preserve non-identifiable attributes. In the audio domain, the Voice Privacy Challenge (Panariello et al., 2024) and the DigitalPhonetics (Meyer et al., 2022) corpus represent the main initiatives for speaker identity anonymization, aiming to preserve both intelligibility and linguistic usability.

Within the framework of multimodality, anonymization has evolved toward systems capable of masking personal information simultaneously across video, text, image, and audio modalities. As described by Shinde et al. (2025), such systems process videos by separating them into two streams—audio and visual frames—which are anonymized independently. The two streams are then recombined to produce a new anonymized video. To the best of our knowledge, no annotated dataset has yet been made publicly available for end-to-end multimodal anonymization.

Altogether, the landscape showcases ongoing progress, while also revealing fragmentation across modalities.[1]

3 Methods

3.1 Dataset compilation

The video corpus was manually compiled by a team of 10 researchers who, leveraging YouTube's filtering tools and diverse search strategies, assembled content that meets predefined criteria for languages, entities, objects, speech, and other general video characteristics. All videos were thoroughly reviewed by the ethical team to ensure the absence of harmful content. Subsequently, the filtered videos were reviewed by the technical team in order to verify that each clip contained clear speech along with identifiable multimodal cues, including faces, voices, and entities (Cook et al., 2019; Baltrušaitis et al., 2019).

The corpus comprises eight European **languages**: Spanish, English, French, Italian, Portuguese, Catalan, Basque, and Galician. It also includes major dialects and regional variants, such as British English, Canadian French, Brazilian Portuguese, and Mexican Spanish.[2]

The total amount of recorded material (approximately 1,756 minutes) was deemed sufficient to obtain statistically reliable metrics for subsequent experiments.

An essential component of an anonymization pipeline is the identification of PII entities. To ensure sufficient examples for evaluating models on PII entity detection, annotators were instructed to select videos containing entities from the following set of target **entities**, either spoken in the audio or visible in the video (e.g., on-screen text or signage): *ADDRESS* (full or partial postal addresses), *AGE* (ages expressed in numbers or words), *DISEASE* (physical or mental illnesses), *EMAIL* (email addresses or media handles), *GPE* (towns, cities, countries, or continents), *ID_NUM* (ID numbers, passports, or license plates), *NAME* (full or partial personal names), *NORP* (nationalities, religions, or affiliations), *NUM_BANK* (account or credit-card numbers), *ORG* (companies, institutions, or associations), *PHONE* (mobile or landline phone numbers), and *ROLE* (professions, positions, or roles).

Furthermore, we include **objects**, which we take to refer to elements within the image that can directly or indirectly reveal a person's identity. Annotators were instructed to select videos containing the object classes such as faces, persons, license plates, logots or screens. A detailed description of these categories can be found in Table 5 (Appendix B).

Most videos feature monologues or short conversations between two people. However, some videos include more than two **speakers**, while others contain no speakers at all (e.g., traffic camera footage without audio, surveillance videos of public spaces). The dataset maintains a balanced distribution of male and female voices and includes a mix of formal and colloquial registers.[3]

To guarantee the thematic variety, 11 different **domains** have been selected from among the most frequent on this platform. The most relevant domains for the corpus—those more likely to contain personal information—are video blogs, where individuals discuss their daily lives, experiences, trips, and so on. However, we also include interviews, news segments, tutorials and other common types of YouTube video, in which other entities and objects may appear, such as account numbers, license

[1] A non-exhaustive list of most promiment tools and projects that make use of datasets compiled for anonymization can be seen in Table 3 in Appendix A

[2] Table 4 in Appendix B presents the distribution of videos and total duration for each language and the full list of dialects.

[3] Figure 2 in Appendix B shows the distribution of videos collected by number of speakers.

plates, and similar information.[4]

In addition to these entities, objects, and domains, compilers were instructed to filter videos that satisfied formal requirements so that all have a Creative Commons Attribution (CC BY) so that derivative work may be distributed, background music accounts for less than 15 percent of the duration, voice overlap should not exceed 15 percent of the total speech duration, video resolution is varied and falls within the 360p to 1920p range, and videos include diverse lighting and ambient conditions (sunny, rainy, etc.).

3.2 Preprocessing

After downlading, the audio tracks were extracted and resampled at 16 kHz. Automatic transcriptions were generated with the open-source ASR system Whisper (Radford et al., 2023). The annotation process was carried out through two parallel workflows: the first focused on transcription, diarization, and the annotation of entities containing private information, and the second on identifying objects within the images.

In Workflow I a team of 15 native speakers representing the eight project languages performed the following tasks.[5] They were tasked with **diarization**, **transcription** and **entity annotation**. Annotators segmented audio into speaker turns, allowing for overlapping segments. To expedite the transcription, annotators were handed automatically generated transcriptions as starting point, which they edited until reaching satisfying quality. Finally, annotators selected audio fragments corresponding to spoken entities within turns and tagged them with one of the 12 predefined entity types. The entity text was carefully transcribed manually, so that transcription errors are reduced to an optimal minimum.

In workflow II, a team of 10 annotators specializing in image annotation performed the following tasks. First they draw rectangular bounding boxes. They labeled objects every 20 frames across the six object classes defined above. Secondly, they labeled objects with attributes: assigning specific visual or semantic properties to objects, such as age, visibility, or accessories for persons; language, type, or context for logos and text; device type for screens; and vehicle type for license plates.

Subsequently, the corpus underwent both automatic and manual review processes. First, the technical team employed unsupervised methods and pretrained models to automatically detect and correct inconsistencies in the classification of objects and entities. Finally, a specialized, independent team of annotators conducted a thorough review of the annotations for all segmentation, classification, and labeling tasks. Due to the expensive nature of the annotation task, annotations were carried out exclusively by a single annotator, thus ruling out the computation of inter-annotator agreement metrics. However, the output was supervised by senior annotators who could correct mislabeling when occurring.

4 Results

The corpus consists of 516 JSON files, each linked to a corresponding YouTube ID, so that they can be identified and downloaded by the corpus users. Besides the annotated speaker turns, entities and objects, each file contains metadata with information about video duration, language, dialect, domain, lighting and atmospheric conditions, and image resolution.

A total of 6,943 entities were annotated in the audio, and the number of objects likely to contain private information exceeds 100,000. The class distribution is presented in Tables 6 and 7 in the Appendix. The Privacy VITA corpus is available upon request to the authors and it will be distributed under a CC-BY-NC-SA license upon publication.

Finally, we provide a preliminary evaluation of the PII identification part of the corpus, comparing an illustrative set of proprietary solutions (LLM-Guard, Presidio and PrivateAI) on the English (EN), Spanish (ES) and French (FR) subsets.[6] The results indicate that current systems still struggle to generalize across languages and entity types. In particular, some entities such as ADDRESS are never identified correctly, and others yet such as ROLE or DISEASE are identified at very low accuracy rates. Furthermore, we observe important differences in performance across languages.

5 Conclusions and future work

Both the scientific community and industry rely on tools and resources that support their research. In the context of anonymization, it is particularly

[4]The distribution of videos across these domains is shown in Figure 1 in Appendix B.

[5]The annotation was carried out using a on-premises deployment of the LabelStudio(Studio, 2026).

[6]Since each model outputs different labels, we normalize them and relabel to match the VITA PII entity set.

Lang	Label	LLM-Guard	Presidio	PrivateAI	N
EN	ADDRESS	0.00	0.00	0.00	46
	AGE	0.00	0.00	0.09	23
	DISEASE	0.00	0.00	0.06	31
	EMAIL	**0.33**	0.29	0.00	5
	GPE	0.58	0.61	0.23	166
	NAME	**0.66**	0.64	0.22	194
	NORP	0.00	**0.55**	0.06	69
	ORG	0.00	0.00	**0.31**	170
	PHONE	0.00	**0.80**	0.00	2
	ROLE	0.00	0.00	0.20	151
	Micro Avg	0.41	0.44	0.19	853
	Macro Avg	0.16	0.29	0.11	853
ES	ADDRESS	0.00	0.00	0.00	36
	AGE	0.00	0.00	**0.70**	27
	DISEASE	0.00	0.00	**0.49**	33
	EMAIL	0.64	0.00	**0.91**	12
	GPE	0.44	0.47	0.50	203
	NAME	**0.52**	0.42	0.51	144
	NORP	0.00	0.00	0.21	66
	ORG	0.00	0.00	**0.46**	162
	PHONE	0.00	0.00	0.00	1
	ROLE	0.00	0.00	0.41	199
	Micro Avg	0.32	0.30	0.36	883
	Macro Avg	0.16	0.09	**0.38**	883
FR	ADDRESS	0.00	0.00	0.00	36
	AGE	0.00	0.00	0.00	4
	DISEASE	0.00	0.00	0.00	17
	EMAIL	0.00	0.00	0.00	5
	GPE	0.23	0.54	0.06	256
	NAME	**0.53**	0.48	0.26	134
	NORP	0.00	0.00	0.24	62
	ORG	0.00	0.00	**0.09**	77
	PHONE	0.00	0.00	0.00	1
	ROLE	0.00	0.00	0.13	51
	Micro Avg	0.25	0.41	0.12	643
	Macro Avg	0.08	0.10	0.07	643

Table 1: Entity-level F1 scores of PII detection systems across languages for a set of proprietary solutions. Support (N) indicates the number of instances per label. Bold indicates the best score for each row.

critical to have a reliable and robust dataset that accounts for the multimodal nature of private data. As the demand for privacy grows, carefully curated and annotated datasets have become an essential resource.

At the same time, developing such tools is challenging, as it demands exceptional care and attention to detail. The task becomes even more complex in a multimodal setting, requiring the coordination of qualified curators across multiple domains, including audio, video, and text. We anticipate that this dataset will showcase the scientific infrastructure supporting the anonymization of personal data and its role within the European ecosystem of ethical and reusable data.

The initial scope of Privacy VITA was limited to specific languages, entities, and objects; however, the corpus is expected to expand both in the number of annotated videos and across additional domains, languages, and entity types.

Additionally, following the approach of Xu et al., 2019, it is important to define and apply quantitative metrics to assess the privacy–utility trade-off. Finally, the pipeline could be enhanced by developing user-friendly interfaces and APIs for real-time applications. Regarding the social impact of our research, such anonymization techniques have the potential to influence open-data policies within the European Union, particularly by shaping the interpretation and implementation of the GDPR. This potential underscores the responsibility to conduct research in a careful, methodical, and transparent manner, ultimately promoting AI as an ethical tool (Floridi and Cowls, 2022; Leslie, 2019).

References

Victoria Arranz, Khalid Choukri, Montse Cuadros, Aitor García-Pablos, Lucie Gianola, Cyril Grouin, Manuel Herranz, Patrick Paroubek, and Pierre Zweigenbaum. 2022. MAPA Project: Ready-to-Go Open-Source Datasets and Deep Learning Technology to Remove Identifying Information from Text Documents. In *Proceedings of the LREC 2022 Joint Workshop on Legal and Ethical Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Language Resources (LEGAL - MDLR 2022)*, pages 64–72, Marseille, France.

Article 29 Data Protection Working Party. 2014. Opinion 05/2014 on anonymisation techniques. Technical report, European Commission.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million facebook profiles harvested for cambridge analytica. *The Guardian*.

Paul Cook, Helena Smith, and Steven Bird. 2019. Multimodal datasets for human communication research: Review and recommendations. *Language Resources and Evaluation*, 53(2):541–565.

Khaled El Emam and Fida K. Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15(5):627–637.

Electronic Privacy Information Center (EPIC). n.d. The privacy act of 1974. `https://epic.org/the-privacy-act-of-1974/`. Accessed: 2025-10-09.

European Parliament and Council. 2016. General data protection regulation (gdpr). *Official Journal of the European Union*, L119:1–88.

Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for ai in society. *Harvard Data Science Review*, 4(1).

IBM Security. 2021. Cost of a data breach report 2021. Technical report, IBM Corporation.

David Leslie. 2019. Understanding artificial intelligence ethics and safety. *CoRR*, abs/1906.05684.

Sarina Meyer, Florian Lux, and Pavel Denisov. 2022. Speaker anonymization with phonetic intermediate representations. In *Interspeech 2022*, pages 4925–4929.

Sarina Meyer, Florian Lux, and Ngoc Thang Vu. 2024. Probing the Feasibility of Multilingual Speaker Anonymization. In *Interspeech 2024*, pages 4448–4452.

Michele Panariello, Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Pierre Champion, Hubert Nourtel, Massimiliano Todisco, Nicholas Evans, Emmanuel Vincent, and Junichi Yamagishi. 2024. The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3477–3491.

AJ Piergiovanni and Michael Ryoo. 2020. Avid dataset: Anonymized videos from diverse countries. *Advances in Neural Information Processing Systems*, 33:16711–16721.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Honolulu, Hawaii.

Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, and Fernando Alonso-Fernandez. 2023. Fiva: facial image and video anonymization and anonymization defense. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 362–371.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Sandip Shinde, Arya Pathak, Shreya Mohite, Sanchitsai Nipanikar, and Keyur Pande. 2025. A multimodal anonymization framework for mp4 videos. In *2025 International Conference on Emerging Smart Computing and Informatics (ESCI)*, pages 1–6.

Label Studio. 2026. Label studio annotation software. Online. [Visited on 2026-03-13].

Verizon Business. 2023. Data breach investigations report 2023. Technical report, Verizon.

Samuel D. Warren and Louis D. Brandeis. 1890. The right to privacy. *Harvard Law Review*, 4(5):193–220.

Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. 2022. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139.

Jie Xu, Feng Zhou, Yong Wang, and Qi Liu. 2019. An evaluation framework for data utility and privacy trade-offs. *IEEE Access*, 7:10372–10383.

A Appendix A: Benchmark Resources

This appendix summarizes representative datasets and tools used in privacy and anonymization research.

A.1 Research datasets

Table 2 provides a compact overview of frequently used resources across modalities and domains.

Dataset / Project	Modality	Data & Privacy scope	Domain
MAPA (Arranz et al., 2022)	Text (multilingual)	Sensitive entities annotated (NER)	Legal / medical
TAB Benchmark (Pilán et al., 2022)	Text (English)	Personal / locational entities labeled	Judicial
Adult (UCI) (Becker and Kohavi, 1996)	Tabular (English)	Structured attributes (age, income, etc.)	Demographics
AViD (Piergiovanni and Ryoo, 2020)	Video (Multiple countries)	Action videos with de-faced subjects	Human activity
PA-HMDB51 (Wu et al., 2022)	Video	Frame-level privacy attributea	Visual privacy
FIVA (Rosberg et al., 2023)	Image / Video	Facial anonymization method (paired data)	Visual
FaceForensics (Rossler et al., 2019)	Video	Real vs. manipulated faces	Deepfake detection
Voice Privacy Challenge (Panariello et al., 2024)	Audio (English)	Speaker identity anonymization	Speech recognition
DigitalPhonetics (Meyer et al., 2022) (Meyer et al., 2024)	Audio (English) (Multilingual)	Voice corpora for anonymization tests	Research / Speech

Table 2: Main datasets and benchmarks related to data anonymization across modalities.

A.2 Tool landscape

Table 3 complements the dataset overview with a practical inventory of available anonymization tools and their modality coverage. This appendix is intended as a quick reference to position Privacy VITA with respect to existing resources and production tools.

B Appendix B: Corpus Composition and Selection Criteria

This appendix reports corpus composition by language, object classes, domains, and number of speakers. Table 4 details the per-language balance in both item counts and recording time.

Table 5 defines the object categories considered during corpus compilation.

Figure 1 summarizes how videos are distributed across thematic domains.

Figure 2 reports the number of speakers per video, which is relevant for diarization and overlap difficulty.

C Appendix C: Annotation Schemas and Counts

This appendix details the annotation taxonomy and the resulting counts for entities and objects. Table 6 reports the observed frequency for each entity type.

Finally, Table 7 summarizes the number of annotated instances for each object class.

Tool	V	I	T	A	No. of Lang.	Entities
Facit	✓		✓		English	Name, e-mail, phone number, address.
Pangeanic (Data Masking)			✓		25+	Name, place, company name, job, family, nationality, age, date, ID, term, event.
Nymiz			✓		102	Address, security verification code, coin, date, IBAN, IP address, location, ID, Social Security number, company name, passport, name, phone number, license plate, website and mail address.
Stick to Digital			✓		Not specified	Name and surname, organization, phone number, address, e-mail, IBAN, numbers, health data.
Presidio			✓		English	Name, location, credit card number, crypto, e-mail, IBAN, IP address, date, Social Security number, phone number, driver license, url, passport, bank number.
ARX			✓		Multilingual	It does not use entity labels, but rather attributes that can be identified as direct, indirect and sensible.
Lingvanex			✓		Spanish	Name, address, +
Delphix			✓		English	Name, address, e-mail, credit card number.
Gallio PRO	✓	✓			NA	Face, license plate.
Eraseid		✓			NA	Face, full body.
ElevenLabs				✓	5+	Dubbing, cloning, changer...
CaseGuard	✓	✓	✓	✓	33+	Face, license plate, screen, people, notebook, vehicle, credit card, PIN number, Social Security number, name, e-mail, phone number, address.

Table 3: Current benchmark (non-exhaustive) of anonymization tools covering V (Video), I (Image), T (Text) and A (Audio) modalities.

Language	No. of videos	Total duration	Dialects
Spanish	72	240 min	Andalusian, Andean, Canarian, Caribbean, Chilean, Central, Mexican, Rioplatense
Catalan	52	162 min	Central, Eastern, Western, Southern
Basque	66	205 min	Central, Labortano, Navarrese, Western
Galician	62	174 min	Central, Eastern, Western
French	67	271 min	Central and Southern African, Northern African, Belgian, Canadian, Standard, Swiss
English	74	270 min	American, Australian, British, Canadian
Italian	56	217 min	Central, Southern, Northern
Portuguese	67	217 min	Brazilian, European
Total	516	1756 min	

Table 4: Number of videos and total duration for each language available on the dataset.

Object class	Definition used during video selection
Faces	Includes partially occluded faces with diverse visual features, racial backgrounds, and ages ranging from minors to the elderly.
Person	Appearing at various depths, in both crowded and solitary scenes, representing diverse ages, races, and visual traits.
License plates	From multiple countries and regions, including the Americas, appearing in cars, motorcycles, vans, and other vehicles with good or poor visibility.
Logos	Typographic or symbolic, with varying visibility and placement.
Texts	Overlaid on images, ranging from signs with personal information (such as directions) to news tickers or crawlers.
Screens	Including computers, laptops, tablets, and smartphone screens.

Table 5: Object classes targeted during corpus compilation and their operational definitions.

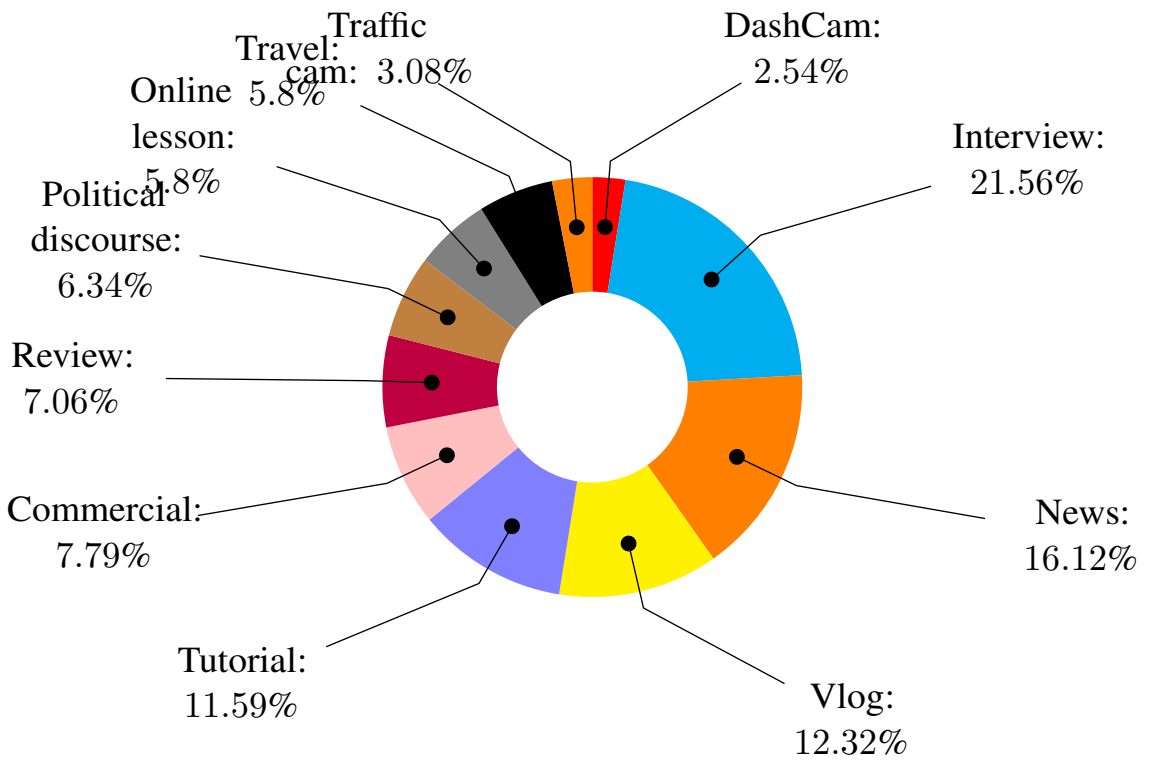

Figure 1: Distribution of domains in the dataset.

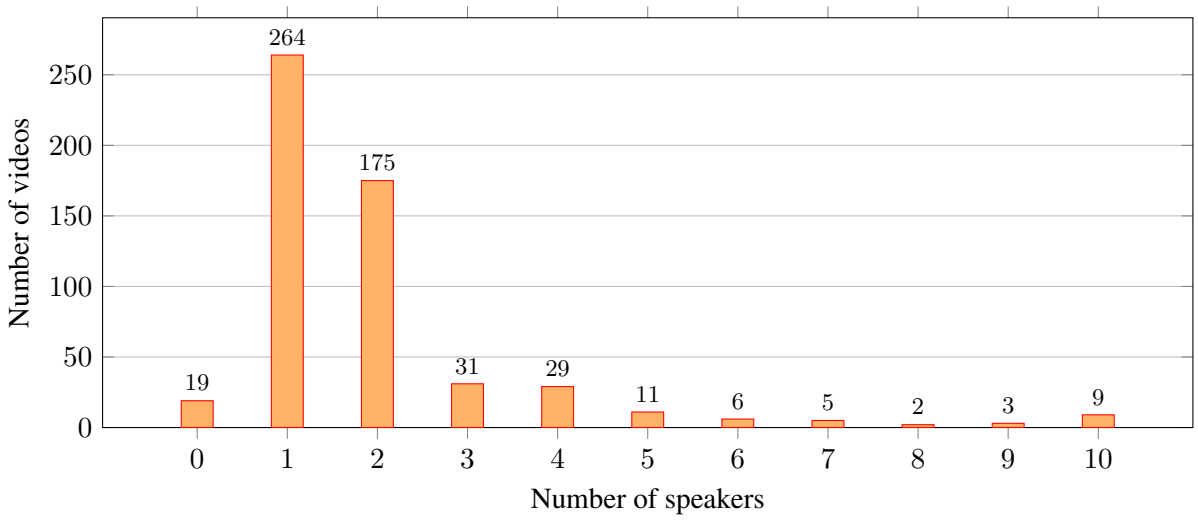

Figure 2: Number of videos by number of speakers.

Entity	Number of examples
GPE	1514
ROLE	1464
ORG	1428
NAME	1196
NORP	546
ADDRESS	264
DISEASE	225
AGE	192
EMAIL	82
PHONE	12
ID_NUM	10
ID_BANK	10
Total	6943

Table 6: Annotated entities in the corpus.

Object type	Number of examples
Face	24875
Person	52100
Plate	850
Logo	19750
Text	10950
Screen	1125
Total	109650

Table 7: Objects found and annotated in the corpus.