shaip

CASE STUDY

Over 30K+ documents web scrapped and annotated into Toxic, Mature, or Sexually Explicit categories

To build automated content moderation Machine Learning Model

COMPANY
A leading Conglomerate

INDUSTRY
Enterprise IT

USE CASE
Automated Content Moderation

OUR OFFERING
- Web Scraping/Data Collection
- Text Classification/Annotation

KEY STATS

Document Collected and classified	No. of Languages	Languages Supported	Project Timeline
30k+ Documents	**2**	**English & Spanish**	**6 months**

OVERVIEW

As social media usage continues to grow, the problem of cyberbullying has surfaced as a significant hurdle for platforms striving to ensure a secure online space. A staggering 38% of individuals encounter this detrimental conduct on a daily basis, emphasizing the urgent demand for inventive content moderation approaches. Organizations today rely on the use of artificial intelligence to address the enduring problem of cyberbullying proactively.

The client was developing robust automated content moderation Machine Learning model for its Cloud offering, for which they were looking for domain-specific vendor who could assist them with accurate training data.

CHALLENGES

The client sought a provider with the necessary expertise and experience in ethically sourcing and annotating large data sets within a specified time period to develop an automated content moderation system using machine learning. The client's essential requirements included:

- Web scraping 15,000 documents in both Spanish and English from prioritized domains
- Categorizing the gathered content into short, medium, and long segments
- Labeling the compiled data as toxic, mature, or sexually explicit content
- Ensuring high-quality annotations with a minimum of 90% accuracy, as measured against a gold standard dataset prepared by the client.

SOLUTION

Leveraging our extensive knowledge in natural language processing (NLP), we assisted the client in gathering, categorizing, and annotating more than 30,000 documents in both English and Spanish. The specifics of the provided solution are as follows:

Data Collection

» Web Scrapping of 15,000 documents each for Spanish and English from the below priority domains

Verticals	Domains
FSI	• Finance • Business and Industry • Government and Laws
Healthcare	• Health and Medical
Manufacturing	• Autos and Vehicles • Computers and electronics • Internet and Communications
Retail	• Retail and shopping • Food and Grocery
Others	• Education and occupation

» Based on the lengths of text, data to be segmented as below

Characters	Examples Include	Segment As	Distribution of Collected Data
50 - 500	Conversational data, social media data	Short Document	20%
501 - 2000	Include email, customer support tickets, complaints, and reviews, including movies, products, restaurants, and book reviews	Medium Document	60%
2001 - 5000	News articles, blogs, online shop product descriptions, publications, system logs, domain-specific documents, Wikipedia articles	Long Document	20%

Text Classification/Annotation

» The documents were classified into the below categories:

Moderation	Document Volume	Description
Adult or Sexually Explicit (SE)	10k	Sexually explicit content/language
Mature	10k	• Adult or Sexually Explicit is mature by default, violence, drugs references, and all content that is not suitable for children of age below 10 yrs. • Abusive to someone with strong language (f*** off, pissed off, and similar).
Toxicity	10k	Rude, disrespectful, or unreasonable language, very hateful, aggressive, insult, identity attack

Examples of Content Moderation

Who the hell Toxic are you? I have never seen an idiot Toxic like you.

I drank too much... I feel so fucked Mature up.

A Bisexual Mature , like a homosexual Mature or a heterosexual Mature , is not defined by sexual activity Mature . (Much like a 15-year-old boy who is attracted to a girl sexually Mature but has never had sex Mature Sexually Explicit and is still straight). A person sexually attracted/aroused Mature Sexually Explicit by the same sex Mature Sexually Explicit and the opposite sex Mature Sexually Explicit is bisexual Mature .

Quality Check and Feedback

In order to satisfy the client's rigorous 90% quality benchmark, Shaip implemented a two-tier quality control process:

» **Level 1: Quality Assurance Check:** 100% of the files to be validated.

» **Level 2: Critical Quality Analysis Check:** Shaips's CQA Team (6⍰ Black Belt holders), who are credentialed and have 10+ years of experience in quality management, will assess the quality of 15%-20% of the retrospective samples.

THE OUTCOME

The training data will help the client build automated content moderation machine learning model that can yield several outcomes that are beneficial for maintaining a safer online environment. Some of the key outcomes include:

Efficiency: It can process vast amounts of data rapidly, reducing the time and effort required to moderate content manually.

Consistency: It can ensure uniform enforcement of content moderation policies across the platform, providing a consistent user experience.

Scalability: It can easily adapt to growing user base and content volumes without the need for significant additional human resources.

Improved Accuracy: It can continuously learn and improve their performance, resulting in more accurate detection of harmful content over time.

Cost-effectiveness: By reducing the reliance on human moderators, automated content moderation can save companies money in terms of labor costs.

Proactive / Real-time Moderation: Machine learning models can identify and remove potentially harmful content as it is generated even before it reaches the broader user base, preventing the negative impact on users.

Customization: Automated content moderation models can be tailored to meet the specific requirements and policies of individual platforms.

However, it is crucial to consider that automated content moderation systems have limitations and might not always be perfect. They can struggle with detecting nuances, sarcasm, and context, and may generate false positives or negatives. It is essential to employ a combination of automated and human moderation to achieve the best results.

CUSTOMER TESTIMONIAL

★★★★★

"We have witnessed a significant improvement in our platform's safety and user experience since implementing an automated content moderation machine learning model. The system's efficiency, consistency, and scalability have allowed us to maintain a secure online environment while accommodating the rapidly growing user base. The machine learning model also adapts to new trends and threats, ensuring that our moderation policies stay up to date."

- Content Moderation Manager, Online Community Platform

shaip

Headquartered in Louisville, Kentucky, Shaip is a fully managed data platform designed for companies looking to solve their most demanding AI challenge enabling smarter, faster, and better results. Shaip supports all aspects of AI training data from data collection, licensing, labeling, transcribing, & de-identifying by scaling seamless of our people, platform, & processes to develop AI/ML models. To learn more about how to make your data science team and leaders' life more manageable, visit us at www.shaip.com.