Subject	Student Information	Date
Research and Project Management in Artificial Intelligence	Last Names: Prida Sánchez	May 26, 2025
	Name: Manuel	

Activity 2: Deployment Platform Proposal for a Hate Speech Detection System on Social Media

1. Introduction and Scope of the Project

1.1. Context and Problem of Hate Speech

The expansion of social media has substantially changed the way people interact and share information, creating a digital environment where opinions circulate rapidly and widely. However, this openness in communication has also brought with it significant challenges. Among these, the spread of hate speech has become a persistent problem in the virtual realm, manifesting itself through expressions that incite contempt or discrimination toward individuals or groups based on their ethnic origin, religious beliefs, sexual orientation, or nationality, among other reasons (United Nations, n.d.). The repercussions of this phenomenon can be serious both on a personal level, where those who are the target of such messages may experience depression, trauma, and an increased risk of suicide (Waqas et al., 2019), and on a collective level, contributing to hostility toward certain groups and encouraging the increase of such messages on social media (Miškolci, 2020).

From a business perspective, the persistence of hate speech can damage the public image of the platforms that host it and reduce user confidence. In light of this situation, the need to effectively detect and manage this type of content has become highly relevant for the scientific community, technology companies, and regulatory bodies. It is therefore essential to have automated solutions that are capable of identifying and mitigating this phenomenon efficiently, and that are accurate and scalable.
This proposal aims to meet this need by seeking to offer professionals responsible for content moderation advanced technological tools, based on artificial intelligence, that facilitate the monitoring of posts on the most widely used social media platforms.

1.2. Scope of the Solution

The purpose of this proposal is to design and implement a deployment platform that supports an artificial intelligence model aimed at detecting hate speech on social media. This platform will be integrated into an architecture capable of receiving real-time data from the APIs of X (formerly Twitter) and Facebook, with the aim of responding immediately to the appearance of harmful content. Once the information has been collected, the messages will undergo a preprocessing process that will

include cleaning, normalization, and tokenization tasks, in order to properly prepare the data for analysis (Haddi et al., 2013). The functional core will consist of an inference system based on the XLM-R model, widely valued for its performance in multilingual text classification (Al-Laith, A., 2025). The results obtained will be stored in a structured database, allowing for subsequent analysis and reporting. To facilitate the work of analysts and moderators, an intuitive graphical interface will be developed to manage and review processed messages, as well as to apply actions such as reporting or deleting posts that violate the rules.

From an operational standpoint, the platform will guarantee continuous availability (24/7) and will have auto-scaling mechanisms that adjust processing capacity based on load, especially in high-activity contexts such as viral events or organized hate campaigns (Mathew et al., 2019). It is a requirement that the response time per message does not exceed two seconds, with the aim of achieving less than one second to ensure a rapid response. In its initial phase, the platform must be capable of handling around 175 messages per second. This figure is an estimate based on current national activity averages: 122.73 tweets per second on X and 52.7 comments per second on Facebook.

Table 1: Estimated tweets per second on X and comments per second on Facebook in Spain. Table created using Statista (2025), Kinsta (2024, 2025), DemandSage (2025), SEO.ai (2025), and Una Vida Online (2024).

Metric	**X (Twitter)**	**Facebook**
Tweets/comments per minute (global)	456	510
Active users (global)	650 million	3,070 million
Active users (Spain)	10.5 million	19.05 million
Proportion of active Spanish users	approx. 1.615%	approx. 0.62%
Tweets/comments per minute in Spain	approx. 7,364	approx. 3,162
Tweets/comments per second in Spain	approx. 122.73	approx. 52.7

Although the platform has been designed to handle a standard load of 175 messages per second, this capacity will need to be expanded to 350 messages per second in exceptional circumstances. This requirement responds to possible peaks in activity generated by viral events, where interaction volumes can double. As a significant precedent, during the 2014 World Cup final, 618,725 tweets per minute were reached, representing a 1.63-fold increase over the average for that year (Yaqub, 2025).

The detection system will be supported by an XLM-R model, composed of some 270 million parameters, enabling it to operate efficiently in multilingual natural language processing tasks (Al-Laith, A., 2025). In addition, it is estimated that up to 2,381 professionals will be able to access the platform simultaneously. This figure reflects the estimated number of staff dedicated to content moderation in Spanish at both X (with 56 employees) and Meta (with 2,325 employees), according to the most recent reports available (X Corp., 2024; Meta Platforms Ireland Limited, 2024).

2. Deployment Platform Architecture

2.1. Hardware and Software Infrastructure

The platform's technical infrastructure will be based on Amazon Web Services (AWS), selected for their reliability, scalability, and the variety of tools they offer for artificial intelligence environments. Priority will be given to Amazon SageMaker, Elastic Compute Cloud (EC2), and Amazon S3, with SageMaker being the main focus in managing the AI model lifecycle. This service will enable both training and parameter adjustment as well as deployment for real-time inference tasks (AWS, 2024a). To meet speed and performance requirements, SageMaker endpoints configured with ml.g4dn.xlarge GPU instances will be used, whose design optimizes the parallel execution of complex operations, a key aspect when dealing with large volumes of real-time data. In parallel, Amazon EC2 will provide the auxiliary computing resources needed to support various platform functions, such as data collection from social media APIs, text preprocessing, and user interface management. Amazon S3 will also be used as a secure and scalable storage system to host the data used in training, the trained model, and the results of message classification. This solution ensures the ability to adapt to the progressive growth in the volume of information derived from social media activity.

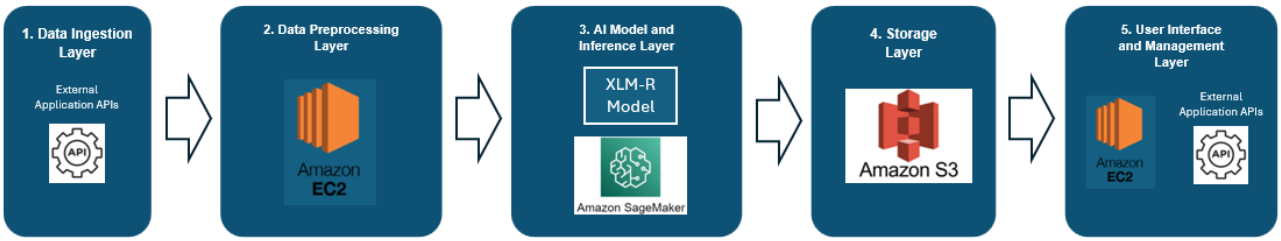

Figure 1: Main Data Flow Architecture. Source: Prepared by the author.

In terms of security, the system design includes advanced measures aimed at protecting the artificial intelligence model and the processed data. To this end, granular permission management is applied using AWS Identity and Access Management (IAM), which is a free service, and follows the principle of least privilege. This access control is complemented by multi-factor authentication (MFA), especially for accounts that handle critical permissions, in order to strengthen protection against unauthorized access. The confidentiality of information is maintained at all times thanks to robust encryption mechanisms. Data stored in Amazon S3 is protected with algorithms such as AES-256, while data in transit is secured using the TLS protocol. In addition, to safeguard data in use, technologies such as AWS Nitro are employed, which allow workloads to be run in hardware-isolated environments at no additional cost as they are included in modern EC2 instances. The compute instances deployed on Amazon EC2 operate within virtual private clouds (VPCs), configured with security rules that act as firewalls to control traffic. These instances receive regular updates to keep the environment free of vulnerabilities. For threat detection and event traceability, the system uses the free AWS CloudTrail basic service, which logs all interactions with AWS services. In addition, specific measures against denial-of-service (DDoS) attacks are incorporated, using the free AWS Shield Standard service. Finally, external access is restricted through IP address whitelists, adding an extra layer of control against unauthorized access.

2.2. Containers and Orchestration

To ensure that the platform is easily portable, scalable, and reproducible, Docker containers will be used. This technology makes it possible to encapsulate both the artificial intelligence model and the supporting applications into individual, self-contained units (Bernstein, 2014). Container management will be handled through Amazon SageMaker Endpoints, a service that enables real-time inference without requiring the user to manage the underlying infrastructure. SageMaker supports the use of custom Docker images as well as preconfigured environments and also provides automatic scaling capabilities based on workload, monitoring parameters such as request volume and latency in order to respond quickly to unexpected traffic spikes. Additionally, the EC2 instances responsible for receiving and processing data also rely on containers, which helps maintain a flexible architecture that is easy to manage.

2.3. Model Lifecycle Management

The AI model lifecycle is divided into the following stages:

Training:

The XLM-R model will be trained on Amazon SageMaker using *ml.g4dn.xlarge* instances, which integrate NVIDIA T4 GPUs and provide an optimal balance between performance and cost for this use case. Training will be carried out every one to two weeks, enabling the continuous incorporation of new data and the progressive refinement of the model's predictive capabilities. Using a pre-trained model such as XLM-R significantly reduces preparation time during this initial phase.

Validation and Testing:

Before the model is deployed to production, it will undergo a thorough validation process. This will include A/B testing to compare different model versions, as well as the evaluation of quantitative metrics such as precision, recall, and F1-score. These metrics will make it possible to objectively assess whether the model's performance is suitable for use in a real-world environment. Ensuring this level of reliability is essential to avoid errors in the classification of sensitive content.

Deployment:

Once the model has successfully passed the testing phase, it will be deployed using SageMaker Endpoints, enabling real-time inference. Auto-scaling mechanisms will be activated to automatically adjust computational resources based on system load, ensuring operational stability even during periods of high demand, such as those triggered by viral events on social media platforms.

Monitoring:

To ensure sustained performance in production, monitoring tools will be enabled to track key aspects in real time, including response latency, prediction quality, and overall service stability. In addition, automated alerting systems will be configured to immediately detect and report any deviations or anomalous behavior, facilitating a rapid response by the technical team.

Retraining and Updating:

To maintain the model's accuracy and relevance in a constantly evolving environment such as social media, periodic retraining will be scheduled. This process will include

the integration of new data and the assessment of potential changes in language trends or expressions of hate speech. Updates will be applied in a controlled manner to ensure a smooth transition between versions and to avoid disruptions to system operation.

2.4. Integration with Other Platforms

The platform architecture is designed to integrate seamlessly with the APIs provided by major social media platforms. This integration enables real-time data ingestion and access to the most up-to-date information. In addition, a proprietary API will be developed to act as a bridge between this solution and external applications, allowing automated access to classification results. Integration with social media APIs is essential for obtaining access to data, while the proprietary API enables the platform's functionality to be reused in other contexts.

3. Cost Estimation and Resource Planning

3.1. Identification of Technological and Human Resources

For the implementation and ongoing operation of the platform, a range of technological and human resources has been identified. From a technological perspective, specific Amazon SageMaker *ml.g4dn.xlarge* instances will be used for both model training and inference. In addition, Amazon EC2 instances will support the platform's underlying infrastructure, while data and model storage will be handled through Amazon S3, leveraging its scalability and reliability. Container orchestration will be managed via Amazon SageMaker Endpoints. Social media platform APIs will also be used for data ingestion.

On the human resources side, the project will rely on specialized roles covering all critical areas. The AI Architect will be responsible for defining the technical direction and establishing the design guidelines for the platform. Model training, validation, and evaluation will be carried out by a Data Scientist, while data ingestion, preprocessing, and storage will be handled by a Data Engineer. An ML/AI Engineer will provide direct support for model deployment and optimization, and at a later stage, an MLOps Engineer will be incorporated to automate deployments and enable continuous monitoring in production. Overall project coordination will be led by an AI Project

Manager, with a Change Manager supporting organizational adoption. Internal and external promotion of the solution will be handled by an AI Evangelist. Additionally, a Business Analyst will define business requirements, a Data Visualization Expert will develop the graphical user interface, and several Frontend and Backend Developers will support the construction of both the interface and the proprietary API. The platform's direct end users will be content analysts and moderators.

3.2. Cost Estimation

The economic viability of the platform for the automatic detection of hate speech is a key pillar for both its initial implementation and its long-term operational sustainability. This section provides a detailed estimate of the monthly costs associated with technological resources, based on the public Amazon Web Services pricing for 2025 in the Ireland region (*eu-west-1*) (AWS, 2025). This region has been selected due to the wider availability of instance types compared to the Spain region (*eu-south-2*). At this preliminary stage of the project, personnel costs are not considered. Currency conversion from US dollars to euros is estimated using the current European Central Bank exchange rate (1.1301). The analysis also incorporates several cost-optimization strategies, including Savings Plans and capacity reservations.

One of the components with the greatest impact on operational costs is the real-time inference system deployed through Amazon SageMaker Endpoints. Under normal conditions, the system is expected to process an average of 197 messages per second. Given that an *ml.g4dn.xlarge* instance can handle approximately 50 messages per second, four continuously active instances would be required. This configuration has a similar cost to using two *ml.g5.xlarge* instances, but it offers greater flexibility if computing capacity needs to be increased more gradually. It is estimated that an *ml.g4dn.xlarge* instance can process between 50 and 70 messages per second when running an XLM-R model, whereas an *ml.g5.xlarge* instance can handle around 100 messages per second. In the *eu-west-1* region, each *ml.g4dn.xlarge* instance costs $0.82 per hour (€0.73/h). This results in an estimated monthly cost of approximately $2,364.48 (€2,091.09) for standard inference alone.

However, during viral events on social media, demand is expected to double, reaching up to 394 messages per second. To handle these activity spikes, the system would need to scale up to eight instances. It is estimated that such situations could occur for

a total of 256 hours per year, divided into 202 hours corresponding to predictable events (listed in Annex 2) and an additional 54 hours attributed to unexpected events such as media controversies, the death of public figures, or spontaneous viral phenomena (see Annex 1). This corresponds to an average of 21.33 hours per month operating in an intensified mode, with an additional monthly cost estimated at approximately $69.96 (€61.88). It should be emphasized that the estimates presented in Annexes 1 and 2 are subjective and that these figures are intended to be indicative only.

Although the on-demand pricing model provides the flexibility required to adapt to workload fluctuations, it is not optimal from a cost-efficiency perspective. For this reason, a hybrid strategy is proposed, combining on-demand pricing with more cost-effective alternatives. Specifically, the use of a three-year Savings Plan is recommended to cover the baseline workload, as it offers a 48% discount in exchange for sustained usage of the *ml.g4dn.xlarge* instance type, resulting in a reduced rate of $0.4256 per hour (€0.3765/h) (AWS, 2024b). This would lower the estimated monthly cost from $2,364.48 (€2,091.09) to $1,225.73 (€1,084.67). To handle traffic spikes, the use of short-term capacity reservations is recommended.

A comparative table of inference costs is presented below:

Table 2: Monthly Inference Costs. Source: Own elaboration

Concept	Cost
Normal Load (Savings Plan)	1,084.67 €
Peak Load (Pay-as-you-go)	61.88 €
Total (Combined Strategy)	**1,146.55 €**

Training is scheduled at a frequency of six sessions per month, each with an estimated duration of three hours. These sessions will run on Amazon SageMaker *ml.g4dn.xlarge* instances, which have an hourly cost of $0.82 (€0.73/h). Under these parameters, the estimated monthly cost for training is approximately $14.76 (€13.06).

Regarding storage, Amazon S3 has been selected. An annual data volume of approximately 2 TB is expected, and to optimize both access and cost, a tiered storage policy will be implemented. Specifically, 500 GB will be kept in the S3 Standard class for the first three months, after which the data will be progressively moved to S3 Glacier Flexible Retrieval, a lower-cost option designed for infrequently accessed files.

Applying current rates of $0.023/GB (€0.0204/GB) for S3 Standard and $0.004/GB (€0.0035/GB) for Glacier (AWS, 2024c), the projected average monthly storage cost for the first year is approximately $12.62 (€11.16).

Table 3: Monthly Amazon S3 Storage Costs for the First Year. Source: Own elaboration

Month	Data in S3 Standard (GB)	S3 Standard Cost ($)	Accumulated Data in S3 Glacier (GB)	S3 Glacier Cost ($)	Total Monthly Cost ($)	Cumulative Cost ($)
1	166.67	$3.67	0	$0.00	$3.67	$3.67
2	333.34	$7.33	0	$0.00	$7.33	$11.00
3	500	$11.00	0	$0.00	$11.00	$22.00
4	500	$11.00	166.67	$0.68	$11.68	$33.68
5	500	$11.00	333.34	$1.35	$12.35	$46.03
6	500	$11.00	500.01	$2.03	$13.03	$59.05
7	500	$11.00	666.68	$2.70	$13.70	$72.75
8	500	$11.00	833.35	$3.38	$14.38	$87.13
9	500	$11.00	1000.02	$4.05	$15.05	$102.18
10	500	$11.00	1166.69	$4.73	$15.73	$117.90
11	500	$11.00	1333.36	$5.40	$16.40	$134.30
12	500	$11.00	1500	$6.08	$17.08	$151.38

For container orchestration, two EC2 t3.medium instances are planned, at a rate of $0.0456 per hour (€0.0403/h), resulting in an estimated monthly cost of $65.66 (€58.09).

Additionally, outgoing data transfer costs should be considered. These are expected to be approximately 5 GB/month (about 3% of 166.67 GB/month). The cost in this case is very low: $0.09/GB (€0.0796/GB), totaling roughly €0.40/month.

The summary of estimated monthly costs is as follows:

Table 4: Summary of Monthly Costs. Source: Own elaboration

Concept	Estimated Monthly Cost
Inference	1,146.55 €
Training	13.06 €
Storage (S3)	11.16 €
Orchestration	58.09 €
Outgoing Data	0.40 €
Total	**1,229.26 €**

This estimate provides a useful starting point for understanding the costs associated with operating the platform. However, it is essential to implement a continuous monitoring system to track usage patterns in order to avoid over-provisioning resources while also identifying potential optimization opportunities. Strategies such as load testing, benchmarking, periodic price reviews, and automated scaling will allow the infrastructure to be dynamically adjusted according to actual demand, thereby preventing unnecessary expenses.

4. Conclusions and Strategic Recommendations

4.1. Justification of the Selected Architecture

The choice of an architecture based on AWS services addresses the need for an infrastructure capable of ensuring scalability, reliability, and high performance—key characteristics for meeting the objectives defined in this project. Specifically, the use of SageMaker Endpoints with GPU instances is justified by their ability to provide very low response times and high performance in real-time inference tasks. Likewise, leveraging pre-trained models such as XLM-R results in significant improvements in efficiency and accuracy, avoiding the computational cost of training from scratch. The inclusion of SageMaker Endpoints as an orchestration solution enables dynamic and agile management of deployments, ensuring system responsiveness even during peak load periods. The combination of these elements creates a robust platform, ready to operate stably in the dynamic and demanding environment of social media.

4.2. Recommendations for Implementation and Sustainability

To ensure the successful implementation of the platform and its long-term maintainability, several strategic recommendations are proposed. First, it is advisable to start with a gradual deployment, launching a pilot version that allows testing the system under real conditions and collecting user feedback from the outset, enabling early adjustments.

Cost control is also essential; therefore, selecting the appropriate resources and optimizing the AI model is crucial to ensure long-term project sustainability. In this regard, Amazon Elastic Inference could be considered in the future as a cost optimization option to achieve greater efficiency in managing inference resources.

A solid data governance framework is also fundamental to ensure that information is handled with quality, security, and respect for privacy. A centralized approach helps reduce potential biases, facilitates audits, and ensures compliance with regulations such as GDPR, while maintaining technical flexibility.

Furthermore, promoting smooth communication and collaboration among developers, operations specialists, and end users is key to ensuring that the solution truly meets their needs. Finally, automating essential tasks—such as deployment, testing, and performance monitoring—through continuous integration tools allows for more efficient workflows, reduces errors, and accelerates improvements.

References

- *Al-Laith, A. (2025, January). Exploring the Effectiveness of Multilingual and Generative Large Language Models for Question Answering in Financial Texts. In Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal) (pp. 230-235).*
- *AWS. (2024a). Amazon SageMaker. Retrieved from https://aws.amazon.com/es/sagemaker/*
- *AWS. (2024b). Savings Plans para machine learning. Retrieved from https://aws.amazon.com/es/savingsplans/ml-pricing/*
- *AWS. (2024c). Amazon S3 pricing. Retrieved from https://aws.amazon.com/es/s3/pricing/*
- *AWS. (2025). AWS Pricing Calculator. https://calculator.aws/#/*
- *Bernstein, D. (2014). Containers and cloud: From lxc to docker to kubernetes. IEEE Cloud Computing, 1(3), 81-84.*
- *DemandSage. (2025). Facebook Users Statistics (2025) — Worldwide Data. Retrieved from https://www.demandsage.com/facebook-statistics/*
- *Haddi, E., Yahioui, A., & Bettayeb, B. (2013). Text pre-processing for information retrieval: A literature review. Recent advances in computer science and information engineering, 45-50.*
- *Kinsta. (2024). Wild and Interesting Facebook Statistics and Facts. Retrieved from https://kinsta.com/blog/facebook-statistics/*
- *Kinsta. (2025). Estadísticas de Twitter: Datos clave sobre la plataforma en 2025. Retrieved from https://kinsta.com/es/blog/estadisticas-twitter/*
- *Mathew, B., Dutt, R., Goyal, P., & Mukherjee, A. (2019, June). Spread of hate speech in online social media. In Proceedings of the 10th ACM conference on web science (pp. 173-182).*
- *Meta Platforms Ireland Limited. (2024, septiembre). Informe de transparencia conforme a la Ley de Servicios Digitales (DSA) – Facebook. Transparency Center. https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebooktransparency.meta.com*

- *Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. Social Science Computer Review, 38(2), 128-146.*
- *SEO.ai. (2025). How Many Users Are on X (Twitter) in 2025?. Retrieved from https://seo.ai/blog/how-many-users-on-x*
- *Statista. (2025). Número de usuarios de Twitter en España. Retrieved from https://es.statista.com/estadisticas/520056/usuarios-de-twitter-en-espana/*
- *Una Vida Online. (2024). Estadísticas del uso de redes sociales en 2024 (España y mundo). Retrieved from https://unavidaonline.com/estadisticas-redes-sociales/*
- *United Nations. (n.d.). Hate speech and the UN strategy and plan of action on hate speech. Retrieved from https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech*
- *Waqas, A., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. PloS one, 14(9), e0222194.*
- *X Corp. (2024, abril). DSA Transparency Report. https://transparency.x.com/dsa-transparency-report.html*
- *Yaqub, M. (2025, enero 31). Twitter Statistics 2025: Growth, Demographics, Usage, and Trends. BusinessDasher. https://www.businessdasher.com/twitter-statistics/*

Appendix 1

Table 5: Estimated Additional Hours on Social Media for Unpredictable Events (Source: Own elaboration)

Event Type	Estimated Events per Year	Average Hours per Event	Social Media Increase (h)
Scandals and Controversies	3	7	21
Deaths of Public Figures	2	9	18
Viral Trends	3	5	15
Weighted Total			**54**

Appendix 2

Table 6: Estimated Additional Hours on Social Media for Planned Events (Source: Own elaboration)

Month	Event	Social Media Increase (h)	Month	Event	Social Media Increase (h)
June 2025	Spanish GP – F1	4	February 2026	Super Bowl	4
	Environment Day	1		NBA All-Star Weekend	3
	Nations League Finals	5		Goya Awards	2
	Wimbledon	4		Grammy Awards	2
	Sónar Festival	2		LaLiga Matches	4
July 2025	Women's Euro Cup	2		Carnival	4
	Tour de France	2		Valentine's Day	4
	San Fermín	6	March 2026	International Women's Day	4
	Mad Cool Festival	2		Oscars	4
August 2025	Arenal Sound	4		LaLiga Matches	4
	Vuelta a España	3		Australian GP – F1	3
	LaLiga Start	3	April 2026	Masters – Augusta	2
	US Open	3		NBA Playoffs Start	3
September 2025	NFL Start	2		LaLiga Matches	4
	LaLiga Matches	4		Copa del Rey Final	5
	World Athletics Championship	1	May 2026	Labor Day	2
October 2025	Hispanic Day	3		Madrid Beach Pro Tour	2
	Mexican GP – F1	4		Champions League Semifinals	4
	LaLiga Matches	3		Eurovision Final	2
	Halloween	3		Europa League Final	4
November 2025	MotoGP Final – Valencia	2		Champions League Final	6
	LaLiga Matches	4		Roland Garros	4
	Brazilian GP – F1	4		LaLiga Matches	4
December 2025	F1 Final – Abu Dhabi	3		GTA VI Launch	4
	Constitution Day	1		Monaco GP – F1	4
	Christmas	5		Cannes Film Festival	3
	LaLiga Matches	3	**Total**		**202**
	New Year's Eve	6			
January 2026	New Year	5			
	Three Kings' Day	4			
	Spanish Super Cup	6			
	LaLiga Matches	4			
	Australian Open	3			