Seguridad del contenido de Azure AI

Qué es la seguridad de contenidos

La seguridad del contenido de Azure AI es un conjunto de características avanzadas de moderación de contenido que se pueden incorporar a sus aplicaciones y servicios. La Seguridad del contenido de Azure AI está disponible como un recurso en Azure Portal.

La protección de contenido en línea es necesaria en un número creciente de situaciones. No solo estamos preocupados por moderar el contenido generado por las personas, sino que también debemos protegerse contra el uso malintencionado de la inteligencia artificial.

Confiar en el contenido generado por el usuario

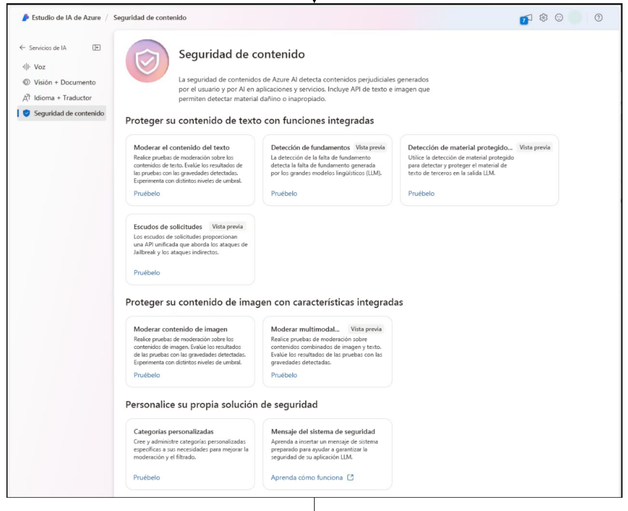
La interacción social es cada vez más parte de muchos espacios digitales. El contenido generado por el usuario original se considera independiente y confiable, y se usa junto con la publicidad y el marketing. Diferentes sectores están animando a sus clientes a conectarse entre sí y su marca.

El contenido dañino tiene muchos efectos negativos. Daña las marcas de confianza, desaconseja a los usuarios de participar en foros en línea y puede tener un impacto devastador en las personas.

La Seguridad del contenido de Azure AI está diseñada para usarse en aplicaciones y servicios para protegerse contra el contenido generado por el usuario y generado por IA perjudicial.

Seguridad del contenido en Azure AI Foundry

La seguridad del contenido de Azure AI está disponible como parte de Azure AI Foundry, una plataforma unificada que le permite explorar muchos servicios de Azure AI diferentes, incluida la seguridad del contenido.



En la página principal de Azure AI Foundry, desplácese hacia abajo y seleccione *Explorar Azure AI Services*. Desde aquí, puede explorar La seguridad del contenido seleccionando **Ver todas las funcionalidades de seguridad de contenido**.

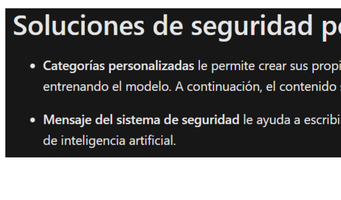
La Seguridad del contenido de Azure AI le permite explorar y probar las características de seguridad de contenido por sí mismo. Seleccione la característica que desea probar y, a continuación, seleccione *Pruébelo*. A continuación, puede usar la interfaz de usuario para probar ejemplos o su propio material. Seleccione *Ver código* para generar código de ejemplo en C#, Java o Python. A continuación, puede copiar y pegar el código de ejemplo y modificar las variables para usar sus propios datos.

¿Cómo funciona la Seguridad del contenido de Azure AI?

La Seguridad del contenido de Azure AI funciona con texto e imágenes y contenido generado por IA.

Las funcionalidades de visión de seguridad de contenido se potencian con el modelo base Florence de Microsoft, que se ha entrenado con miles de millones de pares de imágenes de texto. El análisis de texto usa técnicas de procesamiento de lenguaje natural, lo que proporciona una mejor comprensión de los matices y el contexto. La Seguridad del contenido de Azure AI es multilingüe y puede detectar contenido dañino en formato corto y largo. Actualmente está disponible en inglés, alemán, español, francés, portugués, italiano y chino.

La seguridad del contenido de Azure AI clasifica el contenido en cuatro categorías:



Se usa un nivel de gravedad para cada categoría para determinar si el contenido debe bloquearse, enviarse a un moderador o aprobarse automáticamente.

Las características de seguridad de contenido de Azure AI incluyen:

Protección del contenido de texto

- **Texto moderado** examina texto en cuatro categorías: violencia, odio de voz, contenido sexual y auto-daño. Se devuelve un nivel de gravedad de 0 a 6 para cada categoría. Este nivel ayuda a priorizar lo que necesita atención inmediata por parte de las personas y la urgencia. También puede crear una lista de bloqueados para buscar términos específicos de su situación.
- **Escudos de aviso** es una API unificada para identificar y bloquear ataques de liberación de entradas a LLMs. Incluye tanto la entrada del usuario como los documentos. Estos ataques son avisos a los LLM que intentan omitir las características de seguridad integradas del modelo. Las solicitudes de usuario se prueban para asegurarse de que la entrada en LLM es segura. Los documentos se prueban para asegurarse de que no contienen instrucciones no seguras insertadas en el texto.
- **Detección de materiales protegidos** comprueba el texto generado por IA para texto protegido, como recetas, letras de canciones protegidas u otro material original.
- **Detección de terreno** protege frente a respuestas inexactas en texto generado por IA por LLMs. Los LLM públicos usan los datos disponibles en el momento en que se entrenaron. Sin embargo, los datos se pueden introducir después del entrenamiento original del modelo o se pueden compilar en datos privados. Una respuesta con base es una en la que la salida del modelo se basa en la información de origen. Una respuesta en primer plano es una en la que la salida del modelo varía de la información de origen. La detección de la base incluye una opción de *razonamiento* en la respuesta de la API. Esto agrega un *razonamiento* campo que explica cualquier detección de no en primer plano. Sin embargo, el razonamiento aumenta el tiempo de procesamiento y los costos.

Protección del contenido de la imagen

- **Moderación de imágenes** busca contenido inapropiado en cuatro categorías: violencia, autolesión, sexo y odio. Se devuelve un nivel de gravedad: seguro, bajo o alto. A continuación, establezca un nivel de umbral de bajo, medio o alto. La combinación del nivel de gravedad y umbral determina si la imagen está permitida o bloqueada para cada categoría.
- **Moderación del contenido bidireccional** examina tanto las imágenes como el texto, incluido el texto extraído de una imagen mediante el reconocimiento óptico de caracteres (OCR). El contenido se analiza en cuatro categorías: violencia, odio de voz, contenido sexual y auto-daño.

Soluciones de seguridad personalizadas

- **Categorías personalizadas** le permite crear sus propias categorías proporcionando ejemplos positivos y negativos y entrenando el modelo. A continuación, el contenido se puede examinar según sus propias definiciones de categoría.
- **Mensaje del sistema de seguridad** le ayuda a escribir mensajes eficaces para guiar el comportamiento de un sistema de inteligencia artificial.

Limitaciones

La seguridad del contenido de Azure AI usa algoritmos de inteligencia artificial, por lo que es posible que no siempre detecte un idioma inadecuado. Y, en ocasiones, podría bloquear el lenguaje aceptable porque se basa en algoritmos y aprendizaje automático para detectar lenguaje problemático.

La Seguridad del contenido de Azure AI debe probarse y evaluarse en datos reales antes de implementarse. Y una vez implementado, debe seguir supervisando el sistema para ver con precisión su rendimiento.

Evaluación de la precisión

Al evaluar la precisión de la Seguridad del contenido de Azure AI para su situación, compare su rendimiento con cuatro criterios:

- **Verdadero positivo**: identificación correcta del contenido dañino.
- **Falso positivo**: identificación incorrecta del contenido dañino.
- **Verdadero negativo**: identificación correcta del contenido inofensivo.
- **Falso negativo**: no se identifica el contenido dañino.

La Seguridad del contenido de Azure AI funciona mejor para admitir a moderadores humanos que pueden resolver casos de identificación incorrecta. Cuando las personas agregan contenido a un sitio, no esperan que las publicaciones se quiten sin motivo. Comunicarse con los usuarios sobre por qué el contenido se quita o marca como inapropiado ayuda a todos a comprender lo que está permitido y lo que no es.

Cuándo usar Seguridad del contenido de Azure AI

Muchos sitios en línea animan a los usuarios a compartir sus opiniones. Las personas confían en los comentarios de otras personas sobre productos, servicios, marcas, etc. Estos comentarios suelen ser francos, perspicaces y libres de sesgo de marketing. Pero no todos los contenidos son bienintencionados.

Seguridad del contenido de Azure AI es un servicio de inteligencia artificial diseñado para proporcionar un enfoque más completo para la moderación de contenido. Seguridad del contenido de Azure AI ayuda a las organizaciones a priorizar el trabajo de los moderadores humanos en un número creciente de situaciones:

Education

El número de plataformas de aprendizaje y sitios educativos en línea crece rápidamente, y cada vez se agrega más información. Los educadores deben estar seguros de que los estudiantes no están expuestos a contenidos inapropiados ni están introduciendo solicitudes perjudiciales en los LLM. Además, tanto educadores como estudiantes quieren saber que los contenidos que consumen son correctos y cercanos al material original.

Redes sociales

Las plataformas de redes sociales son dinámicas y rápidas, lo que requiere moderación en tiempo real. La moderación del contenido generado por el usuario incluye publicaciones, comentarios e imágenes. Seguridad del contenido de Azure AI ayuda a moderar los contenidos matizados y multilingües para identificar el material dañino.

Marcas

Las marcas usan cada vez más las salas de chat y los foros de mensajes para animar a sus clientes fieles a compartir sus opiniones. Sin embargo, el material ofensivo puede dañar una marca y evitar que los clientes contribuyan. Quieren tener la seguridad de que el material inapropiado puede identificarse y eliminarse rápidamente. Las marcas también están agregando servicios de IA generativa para ayudar a las personas a comunicarse con ellas y, por lo tanto, necesitan protegerse de los malos actores que intentan explotar los modelos de lenguaje grandes (LLM).

Comercio electrónico

Los contenidos de los usuarios se generan reseñando productos y comentándolos con otras personas. Este material es un poderoso instrumento de marketing, pero cuando se publican contenidos inapropiados se daña la confianza de los consumidores. Además, los problemas normativos y de cumplimiento son cada vez más importantes. Seguridad del contenido de Azure AI ayuda a filtrar los listados de productos en busca de reseñas falsas y otros contenidos no deseados.

Juegos

Los juegos son un ámbito difícil de moderar debido a sus gráficos muy visuales y a menudo violentos. Los juegos tienen comunidades fuertes en las que la gente comparte con entusiasmo sus progresos y experiencias. La compatibilidad con moderadores humanos para mantener la seguridad en los juegos incluye la supervisión de avatares, nombres de usuario, imágenes y materiales basados en texto. Seguridad del contenido de Azure AI cuenta con herramientas avanzadas de visión de IA para ayudar a las plataformas de juegos moderados a detectar conductas indebidas.

Servicios de IA generativa

Las organizaciones usan cada vez más servicios de IA generativa para permitir el acceso a datos internos más fácilmente. Para mantener la integridad y la seguridad de los datos internos, se deben comprobar las indicaciones del usuario y las salidas generadas por la IA para evitar el uso malintencionado de estos sistemas.

Noticias

Los sitios web de noticias deben moderar los comentarios de los usuarios para evitar la propagación de información errónea. Seguridad del contenido de Azure AI puede identificar lenguaje que incluya incitación al odio y otros contenidos dañinos.

Otras situaciones

Hay muchas otras situaciones en las que el contenido debe moderarse. Seguridad del contenido de Azure AI se puede personalizar para identificar el lenguaje problemático para casos específicos.