Sistema de Recomendación de Películas Basado en Análisis de Sentimientos

Iván Mauricio Melo

Universidad Autónoma de Occidente

Santiago de Cali, Colombia

ivan.melo@uao.edu.co

Johan Sebastián Tobar

Universidad Autónoma de Occidente

Santiago de Cali, Colombia

Johan.patiño@uao.edu.co

I. INTRODUCCIÓN.

En la actualidad, el acceso a contenido digital, particularmente en plataformas de streaming, ha transformado la forma en que las personas consumen entretenimiento. Con un catálogo que incluye miles de películas y series, el desafío ya no es encontrar contenido, sino descubrir qué es lo más relevante y atractivo para cada usuario. Este problema ha dado lugar a la evolución de sistemas de recomendación, herramientas clave para filtrar información y personalizar experiencias. Sin embargo, muchos de estos sistemas tradicionales basan sus sugerencias en datos explícitos como calificaciones, géneros o historial de visualización, dejando de lado un factor fundamental: las emociones humanas.

Las emociones y sentimientos expresados en las reseñas de los usuarios son un recurso valioso, ya que capturan aspectos subjetivos que las calificaciones numéricas no pueden representar. Una película puede ser "dramática", "inspiradora" o "emocionalmente intensa", y estas características pueden ser determinantes en la decisión de un usuario para verla. Ignorar estas dimensiones emocionales resulta en recomendaciones genéricas que no logran conectar profundamente con las preferencias individuales.

Este proyecto aborda esta problemática proponiendo un sistema de recomendación de películas innovador, capaz de analizar y extraer las emociones expresadas en las reseñas de usuarios. A través de técnicas avanzadas de aprendizaje automático, como embeddings semánticos y análisis de sentimientos, el sistema identifica patrones emocionales y los combina con características contextuales, logrando sugerencias personalizadas.

Además, este proyecto integra técnicas de reducción de dimensionalidad, como UMAP, para visualizar las relaciones entre películas y ofrecer una representación gráfica de cómo las emociones y características contextuales influyen en las conexiones entre títulos. Esta visualización no solo mejora la comprensión del modelo, sino que también ayuda a los usuarios a explorar recomendaciones de manera más intuitiva.

En un contexto donde la personalización y la experiencia del usuario son clave para la competitividad de las plataformas digitales, este sistema busca no solo mejorar la calidad de las recomendaciones, sino también redefinir la forma en que las emociones se integran en la toma de decisiones. Este enfoque, que combina análisis de sentimientos, aprendizaje automático y visualización de datos, representa un avance significativo en la personalización de contenido en la industria del entretenimiento.

II. MARCO TEORICO.

Sistemas de Recomendación

Los sistemas de recomendación son algoritmos diseñados para predecir las preferencias de los usuarios y recomendar productos o servicios que puedan ser de su interés. Estos sistemas están presentes en muchas plataformas de streaming, tiendas online y redes sociales, entre otros. Existen tres enfoques principales:

Sistemas Basados en Contenido: Este enfoque utiliza características explícitas del contenido, como género, actores, directores, o incluso descripciones de las películas, para realizar recomendaciones. Los algoritmos comparan las características de los elementos que un usuario ha consumido previamente con otras disponibles en el sistema. La principal limitación de este enfoque es la "falta de novedad" o la incapacidad de recomendar productos que son significativamente diferentes de los que el usuario ya ha visto.

Sistemas Basados en Colaboración: Estos sistemas se basan en las interacciones de los usuarios con el contenido, como calificaciones o clics, y buscan similitudes entre los comportamientos de diferentes usuarios. Si dos usuarios tienen preferencias similares en el pasado, se les recomendarán elementos que uno de ellos haya disfrutado. Este enfoque tiene la ventaja de poder recomendar productos no similares, pero su principal inconveniente es el problema de "arranque en frío", es decir, la dificultad de realizar recomendaciones cuando hay poca información sobre los usuarios o los elementos nuevos.

Sistemas Híbridos: Combina los enfoques anteriores para aprovechar sus ventajas y mitigar sus limitaciones. Los sistemas híbridos pueden combinar filtrado colaborativo y basado en contenido, mejorando la precisión y la capacidad de recomendar elementos nuevos o inexplorados.

2.2. Análisis de Sentimientos

El análisis de sentimientos es una rama del procesamiento de lenguaje natural (NLP) que se encarga de identificar y extraer opiniones, emociones o actitudes subyacentes en un texto. En el contexto de reseñas de películas, el análisis de sentimientos permite determinar si un usuario tiene una visión positiva, negativa o neutral sobre una película. Este análisis es crucial porque las emociones y sentimientos pueden influir profundamente en las decisiones de los usuarios, y capturarlos puede proporcionar una perspectiva más precisa sobre sus preferencias.

El análisis de sentimientos se puede abordar utilizando técnicas de aprendizaje automático o aprendizaje profundo. En este proyecto, se utiliza un modelo basado en SentenceTransformers, que genera representaciones vectoriales (embeddings) del texto. Los embeddings son una forma compacta y eficiente de representar un texto en un espacio de alta dimensión, lo que permite comparar y analizar la similitud semántica entre diferentes textos.

2.3. Embeddings Semánticos

El concepto de embeddings semánticos se refiere a la representación de objetos (como palabras, frases o documentos) en un espacio vectorial en el que la distancia entre los vectores refleja las similitudes semánticas entre ellos. Por ejemplo, dos películas que comparten una temática similar o emociones similares deberían tener embeddings cercanos en el espacio vectorial.

Uno de los modelos más populares para generar estos embeddings es SentenceTransformer, que es capaz de mapear textos de longitud variable a un espacio de vectores de dimensión fija. Este modelo aprovecha Transformers, como BERT (Bidirectional Encoder Representations from Transformers), para capturar las relaciones semánticas y contextuales dentro de las oraciones. Al usar embeddings generados con SentenceTransformer, se obtiene una representación densa y significativa de las reseñas, lo que permite comparar la similitud entre películas a nivel emocional y temático.

2.4. Reducción de Dimensionalidad

La reducción de dimensionalidad es una técnica utilizada para reducir la cantidad de variables en un conjunto de datos, conservando la mayor cantidad posible

de información relevante. Esta técnica es particularmente útil cuando se trabaja con embeddings de alta dimensión, como los generados por modelos de lenguaje como SentenceTransformer.

Uno de los métodos más utilizados para la reducción de dimensionalidad es el UMAP (Uniform Manifold Approximation and Projection). UMAP es una técnica no lineal que conserva la estructura local y global de los datos, lo que significa que las relaciones entre los puntos de datos en el espacio original se mantienen al reducir la dimensionalidad. En este proyecto, UMAP se utiliza para representar las similitudes emocionales y temáticas entre las películas en un espacio bidimensional. Esta visualización facilita la comprensión de las relaciones entre las películas y ayuda a detectar patrones, como la agrupación de películas por género o sentimiento.

2.5. Visualización de Datos en Sistemas de Recomendación

La visualización de datos es fundamental en los sistemas de recomendación, ya que permite a los usuarios o desarrolladores interpretar y analizar las relaciones entre los elementos recomendados. A través de técnicas como UMAP, es posible visualizar cómo las películas se agrupan según sus características emocionales y semánticas, lo que mejora la interpretación de las recomendaciones.

Al proyectar los embeddings de las películas en un espacio bidimensional, los desarrolladores pueden observar cómo las películas con emociones similares tienden a agruparse, y cómo las características de género o tema afectan la disposición espacial de los elementos. Esto no solo ayuda a evaluar la calidad del sistema de recomendación, sino que también proporciona una visión más intuitiva de cómo se construyen las recomendaciones.

2.6. Relevancia del Análisis de Sentimientos en Recomendaciones Personalizadas

El análisis de sentimientos, cuando se aplica a las reseñas de los usuarios, proporciona una dimensión adicional a los sistemas de recomendación. En lugar de basarse únicamente en datos objetivos como las calificaciones o las características explícitas de las películas, este enfoque considera las emociones y opiniones subjetivas de los usuarios. Esto permite a los sistemas ofrecer recomendaciones más personalizadas, ya que las emociones de un usuario hacia una película pueden reflejar mejor sus verdaderas preferencias que las categorías tradicionales.

Además, integrar análisis de sentimientos puede mejorar la capacidad del sistema para lidiar con películas con géneros híbridos o difíciles de clasificar de manera convencional, ya que las emociones que generan pueden ser más consistentes que las etiquetas de género.

III. DESCRIPCION DEL PROBLEMA.

En la actualidad, los usuarios de plataformas de streaming, como Netflix, Amazon Prime Video y otras, se enfrentan a un vasto mar de opciones de contenido. Este exceso de alternativas genera una sobrecarga cognitiva, donde decidir qué película o serie ver puede resultar abrumador. Las plataformas tradicionales han intentado abordar este problema utilizando sistemas de recomendación que generalmente se basan en dos enfoques principales:

- Recomendaciones basadas en contenido: Este enfoque se basa en las características explícitas de las películas, como el género, el director o los actores. Aunque útil, este sistema no tiene en cuenta las emociones o la subjetividad detrás de las reseñas de los usuarios, lo que puede llevar a sugerencias que no siempre se alinean con las preferencias personales de los usuarios.

- Recomendaciones colaborativas: Este enfoque se basa en el comportamiento de otros usuarios, recomendando elementos que han sido valorados positivamente por personas con gustos similares. Sin

embargo, los sistemas colaborativos sufren del problema de la sparsity, es decir, que no pueden generar recomendaciones precisas para nuevos usuarios (falta de historial) o para películas que no tienen una amplia base de usuarios.

Ambos enfoques, por separado, no logran capturar la complejidad emocional que los usuarios experimentan al elegir contenido. Por ejemplo, una persona podría estar interesada en una película romántica en un día, pero prefiera algo de acción al siguiente. Las reseñas de los usuarios a menudo contienen pistas valiosas sobre estas emociones y preferencias subjetivas que no son fácilmente capturadas por los sistemas convencionales.

El Desafío de la Subjetividad en las Preferencias de los Usuarios

Las decisiones sobre qué contenido ver no siempre se basan únicamente en características objetivas como el género o las calificaciones. De hecho, las emociones juegan un papel crucial en la elección de películas. Por ejemplo, un usuario puede estar buscando una película para relajarse, sentirse inspirado o experimentar una montaña rusa emocional, lo cual está fuertemente influenciado por el análisis de sentimientos que las reseñas pueden revelar.

Sin embargo, los sistemas de recomendación tradicionales no capturan adecuadamente estos matices emocionales, y solo se enfocan en datos más objetivos. Este problema limita la capacidad de los sistemas de recomendación para ofrecer experiencias verdaderamente personalizadas, lo que puede llevar a la insatisfacción del usuario y una menor retención en las plataformas.

La Oportunidad para Mejorar

La incorporación de análisis de sentimientos a los sistemas de recomendación presenta una solución innovadora a este problema. El análisis de sentimientos permite extraer la emoción subyacente en las reseñas de los usuarios, lo que puede proporcionar una mejor comprensión de las verdaderas preferencias del usuario. Al integrar estos análisis en los sistemas de recomendación, se pueden identificar patrones emocionales que están estrechamente relacionados con las elecciones de contenido, lo que resulta en sugerencias más personalizadas.

Al aprovechar las emociones expresadas en las reseñas, este sistema de recomendación puede ir más allá de los enfoques convencionales, proporcionando recomendaciones basadas no solo en las características objetivas del contenido, sino también en las respuestas emocionales que estas películas generan en los espectadores. Esta mejora en la personalización tiene el potencial de mejorar significativamente la experiencia del usuario, aumentando la satisfacción y fidelidad a largo plazo con la plataforma de streaming.

Objetivo General

- Desarrollar un sistema de recomendación de películas que utilice análisis de sentimientos y técnicas de representación semántica para mejorar la personalización y precisión de las recomendaciones.

Objetivos Específicos

- Implementar un modelo de análisis de sentimientos para extraer emociones de las reseñas de usuarios. Utilizar técnicas de reducción de dimensionalidad para visualizar las relaciones entre películas según sus similitudes emocionales.

IV.JUSTIFICACION DEL PROYECTO

La creación de un sistema de recomendación basado en el análisis de

sentimientos es una respuesta innovadora a la creciente demanda de personalización y relevancia en las recomendaciones de contenido digital, especialmente en plataformas de streaming de películas. Este proyecto no solo aborda una necesidad tecnológica, sino también un vacío en la forma en que los sistemas tradicionales abordan las preferencias de los usuarios. A continuación se explican las razones clave que sustentan la importancia y relevancia de este enfoque:

5.1. Mejora de la Experiencia del Usuario

Los sistemas tradicionales de recomendación, que se basan principalmente en algoritmos de filtrado colaborativo o en características explícitas de los contenidos (como el género o las calificaciones), tienen limitaciones cuando se trata de capturar las sutilezas emocionales y subjetivas de las preferencias de los usuarios. Al integrar el análisis de sentimientos, el sistema puede comprender no solo qué tipo de películas les gustan a los usuarios, sino también cómo se sienten acerca de ellas, lo cual es crucial para una experiencia más rica y personalizada.

Este enfoque proporciona recomendaciones más profundas y matizadas, al reflejar no solo las preferencias tradicionales (por ejemplo, si a un usuario le gustan las películas de acción), sino también sus respuestas emocionales ante las historias o temas tratados en las películas. Esto no solo mejora la experiencia de navegación, sino que también aumenta la satisfacción general del usuario, permitiendo que las recomendaciones se alineen mejor con su estado de ánimo o emociones actuales.

5.2. Respuesta a la Necesidad de Personalización Avanzada

En un mercado donde la competencia entre plataformas de streaming es feroz, las empresas necesitan herramientas más efectivas para fidelizar a los usuarios y ofrecerles contenido que realmente resuene con sus emociones y gustos. Los usuarios actuales demandan experiencias cada vez más personalizadas, donde las recomendaciones no solo estén basadas en patrones estadísticos, sino que reflejen aspectos más humanos y subjetivos de la preferencia individual.

El análisis de sentimientos, combinado con las representaciones semánticas de las reseñas, permite una personalización más profunda, al identificar patrones emocionales y temáticos que van más allá de las clasificaciones convencionales. Esto se traduce en un mayor grado de satisfacción y un uso más prolongado de la plataforma, lo cual tiene un impacto directo en las métricas de retención y lealtad de los usuarios.

5.3. Innovación en la Industria del Entretenimiento

Este proyecto introduce un enfoque innovador que fusiona el análisis de sentimientos con la recomendación de contenido en la industria del entretenimiento. La integración de técnicas avanzadas de procesamiento de lenguaje natural (PLN), como los embeddings generados por modelos de deep learning (en este caso, SentenceTransformer), aporta una nueva dimensión a la forma en que las plataformas pueden interactuar con sus usuarios.

La incorporación de UMAP para la visualización de las relaciones emocionales y contextuales entre películas también ofrece una nueva perspectiva sobre cómo analizar y presentar los datos de películas de manera que resuene mejor con el espectador. Estas técnicas podrían ser adaptadas a otros campos dentro de la industria del entretenimiento, como la música, los libros o incluso los videojuegos, lo que abre nuevas posibilidades para la personalización de servicios en línea.

5.4. Relevancia Social y Comercial

El sistema de recomendación propuesto tiene el potencial de mejorar no solo la experiencia individual del usuario, sino también la interacción social dentro de las plataformas de streaming. A medida que los usuarios reciben recomendaciones más precisas y alineadas con sus emociones, podrían compartir sus experiencias de forma más activa, lo que fomentaría una comunidad más comprometida.

Desde el punto de vista comercial, las plataformas de streaming que adopten este enfoque innovador pueden diferenciarse de sus competidores al ofrecer una experiencia única, con recomendaciones más acertadas, lo que podría traducirse en una ventaja competitiva en el mercado. Además, la capacidad de generar recomendaciones emocionales también abre la puerta a nuevas formas de publicidad y monetización, mediante la personalización de anuncios según el estado de ánimo o preferencias emocionales de los usuarios.

5.5. Posibilidad de Mejoras Continuas y Futuras Investigaciones

El sistema no solo es relevante en su forma actual, sino que también abre nuevas oportunidades para la investigación y el desarrollo futuro. A medida que se recogen más datos de usuarios, el modelo puede refinarse mediante técnicas de retroalimentación continua, mejorando la precisión de las recomendaciones y adaptándose a nuevas tendencias emocionales. Este enfoque iterativo no solo es un avance en la tecnología de recomendación, sino que también demuestra la capacidad de aprendizaje adaptativo de los modelos de IA en tiempo real.

5.6. Impacto en la Retención de Usuarios

En la actualidad, la retención de usuarios en plataformas de streaming es uno de los principales desafíos de las empresas del sector. Los sistemas de recomendación inteligentes que ofrecen contenido relevante basado en el estado emocional de los usuarios pueden aumentar significativamente el tiempo de permanencia en la plataforma y la tasa de retorno de los usuarios. Un sistema que se adapta a los cambios emocionales de los usuarios, sugiriendo contenido en función de sus sentimientos y preferencias cambiantes, es más probable que mantenga su interés y los motive a volver.

V. PLANTEAMIENTO DE LA SOLUCION

El sistema de recomendación propuesto integra una serie de técnicas avanzadas para mejorar la personalización y la precisión en las recomendaciones de películas, basándose en el análisis de sentimientos de las reseñas de los usuarios. La solución sigue un enfoque híbrido que combina varias metodologías, desde el análisis de sentimientos hasta la representación semántica y la visualización de relaciones. La estructura de la solución es la siguiente:

6.1. Análisis de Sentimientos

El primer componente del sistema es el análisis de sentimientos, que se centra en extraer las emociones y opiniones expresadas en las reseñas de los usuarios. Este proceso se realiza utilizando un modelo de análisis de sentimientos preentrenado, que clasifica las reseñas en categorías emocionales (por ejemplo, positiva, negativa, neutral) y extrae temas y palabras clave asociadas a cada sentimiento. Esta capa de análisis de sentimientos permite que el sistema entienda el tono emocional de las reseñas, lo que ayuda a inferir las preferencias y estados de ánimo de los usuarios, aspectos que tradicionalmente no se capturan en sistemas de recomendación basados en características o calificaciones numéricas.

6.2. Embeddings Semánticos

El siguiente componente consiste en representar las películas y sus reseñas en un espacio vectorial utilizando embeddings semánticos. Los embeddings son representaciones numéricas de objetos (en este caso, películas) que capturan la semántica y el contexto de los datos. En el caso de este sistema, se utiliza el modelo preentrenado all-MiniLM-L6-v2 de SentenceTransformers, que genera representaciones compactas y precisas de las reseñas de las películas.

Estos embeddings son útiles porque permiten que las películas sean representadas en un espacio vectorial donde las relaciones de similitud entre ellas se reflejan de forma matemática. Las películas con reseñas similares o con emociones compartidas estarán más cerca en este espacio, mientras que las películas con reseñas significativamente diferentes se alejarán. Esto es crucial para generar recomendaciones que no solo consideren características explícitas de las películas (como el género o la calificación), sino también el

contexto emocional y subjetivo de las reseñas de los usuarios.

6.3. Reducción de Dimensionalidad con UMAP

Para visualizar y comprender mejor las relaciones entre las películas representadas en el espacio vectorial, se utiliza la técnica de reducción de dimensionalidad llamada UMAP (Uniform Manifold Approximation and Projection). UMAP toma los embeddings generados por el modelo de SentenceTransformers y los proyecta en un espacio de dos dimensiones. Este paso facilita la visualización de cómo se agrupan las películas según sus similitudes emocionales, temáticas y contextuales.

La visualización de las películas en un espacio bidimensional no solo proporciona una forma de explorar los datos, sino que también permite identificar patrones y agrupamientos de películas que comparten emociones o géneros similares. Este enfoque ayuda a los desarrolladores a comprender la distribución de las películas en el espacio de características y a mejorar la precisión de las recomendaciones.

6.4. Sistema de Recomendación

El sistema de recomendación final utiliza las representaciones vectoriales de las películas (provenientes de los embeddings semánticos) y las visualizaciones generadas por UMAP para realizar sugerencias personalizadas. El proceso de recomendación se basa en calcular las similitudes entre las películas en el espacio de embeddings. Cuando un usuario ingresa una película que le gusta, el sistema puede encontrar otras películas que se encuentren cercanas en el espacio vectorial, es decir, aquellas que comparten emociones y temas similares.

El sistema puede ajustar la relevancia de las recomendaciones utilizando tanto el análisis de sentimientos como las características del contenido, lo que da como resultado recomendaciones más matizadas y contextualizadas. Por ejemplo, si un usuario disfruta de una película con una reseña positiva llena de emoción y esperanza, el sistema podrá sugerir otras películas que compartan sentimientos similares en sus reseñas, proporcionando una experiencia más personalizada que los sistemas tradicionales basados únicamente en géneros o calificaciones.

6.5. Interacción de Componentes

El sistema funciona de manera híbrida, donde el análisis de sentimientos y la reducción de dimensionalidad son complementarios. La interacción de estos componentes permite:

La clasificación emocional de las reseñas, que enriquece las recomendaciones, no solo basadas en las características explícitas de las películas, sino también en el tono emocional subyacente.

La proyección de los embeddings de las películas en un espacio bidimensional, lo que facilita la comprensión de sus relaciones semánticas y la identificación de agrupamientos significativos.

La recomendación de películas que son semánticamente similares, lo que hace que las sugerencias sean más relevantes para el usuario, considerando tanto los aspectos emocionales como los temáticos.

Este enfoque híbrido proporciona una experiencia de usuario más rica y dinámica, ya que tiene en cuenta la complejidad de las preferencias humanas, que van más allá de las simples calificaciones numéricas o categorías de género.

6.6. Impacto y Aplicabilidad

El sistema de recomendación propuesto tiene un impacto significativo en la industria del entretenimiento, ya que ofrece una experiencia más personalizada y emocionalmente inteligente. Este enfoque innovador puede ser adaptado no solo para películas, sino también para otros tipos de contenido, como series de televisión, música o libros, siempre que exista una base de datos de reseñas de usuarios.

VI. JUSTIFICACION DEL DATA SET

La elección del dataset IMDb Top 1000 de Kaggle para este proyecto de recomendación de películas está fundamentada en varias razones clave que aseguran la pertinencia y utilidad del conjunto de datos en el contexto de

análisis de sentimientos y sistemas de recomendación. A continuación, se detallan los aspectos más relevantes que justifican esta elección:

1. Calidad y Variedad de Datos

El dataset IMDb Top 1000 contiene información crucial y detallada sobre las películas, tales como:

- **Título de la película:** Esencial para la identificación de cada película en el sistema.
- **Género:** Permite categorizar las películas en distintos géneros, lo que es útil para realizar recomendaciones basadas en las preferencias de los usuarios por ciertos tipos de contenido.
- **Año de estreno:** Es un factor relevante para los usuarios al elegir películas según la época.
- **Duración:** Indica la longitud de las películas, lo que también puede ser un criterio para algunos usuarios a la hora de decidir qué ver.
- **Clasificación de IMDb y Metascore:** Estas métricas de popularidad y calidad permiten integrar un análisis más amplio de la película, lo cual puede ser utilizado para ponderar las recomendaciones.
- **Reseñas de usuarios:** Este es uno de los aspectos más valiosos del dataset, ya que las reseñas pueden ser analizadas para extraer las emociones y opiniones de los usuarios, un elemento fundamental para el análisis de sentimientos.

2. Enfoque Específico en Películas Populares

El dataset está basado en las 1000 películas mejor clasificadas en IMDb, lo que garantiza que las películas en el conjunto de datos son populares y bien valoradas por una amplia base de usuarios. Esto permite centrarse en contenido con alta aceptación y más probabilidades de atraer a una audiencia general, lo que aumenta la relevancia de las recomendaciones generadas por el sistema.

3. Enfoque en las Preferencias de los Usuarios

Al incluir calificaciones y reseñas, el dataset facilita la extracción de información emocional y subjetiva. Este aspecto es clave para el análisis de sentimientos, ya que no solo se cuenta con las valoraciones objetivas (como las calificaciones numéricas), sino también con la interpretación emocional contenida en las reseñas. A través de técnicas de procesamiento de lenguaje natural (NLP), estas reseñas se pueden analizar para obtener una comprensión más profunda de lo que los usuarios valoran o rechazan de una película.

4. Diversidad de Géneros y Estilos

Aunque el dataset está compuesto por películas de alto rendimiento en IMDb, estas películas abarcan una gran diversidad de géneros, desde dramas, comedias, thrillers hasta documentales y películas de ciencia ficción. Esta diversidad permite que el sistema de recomendación no se limite a un solo tipo de contenido, sino que pueda sugerir una amplia gama de películas basadas en las emociones o características semánticas que los usuarios prefieren. Además, la heterogeneidad de géneros hace posible la creación de un sistema más dinámico y flexible que pueda adaptarse a diferentes tipos de usuarios.

5. Facilidad para Implementación de Técnicas Avanzadas

El dataset está bien estructurado y es fácil de integrar con técnicas avanzadas de procesamiento de datos, como:

Análisis de sentimientos: Las reseñas de los usuarios proporcionan un contexto perfecto para aplicar técnicas de análisis de sentimientos, lo que permitirá identificar las emociones predominantes en las opiniones de los usuarios.

Reducción de dimensionalidad: La diversidad de características, como los géneros, el año de estreno, y las clasificaciones, permite aplicar técnicas de reducción de dimensionalidad (como UMAP) para visualizar relaciones complejas entre las películas de manera más comprensible.

Embeddings semánticos: Las descripciones textuales y las reseñas de los usuarios pueden ser convertidas en embeddings semánticos, lo que permite encontrar patrones y similitudes emocionales y contextuales entre las películas.

6. Popularidad y Reproducibilidad

El dataset IMDb Top 1000 es ampliamente utilizado y bien conocido en la comunidad de investigación, lo que facilita la comparabilidad con otros estudios y la reproducibilidad de los resultados. Además, es accesible públicamente a través de Kaggle, lo que permite que otros investigadores y desarrolladores reproduzcan el análisis y extiendan la investigación si lo desean. Esta accesibilidad mejora la transparencia del proyecto y facilita futuras contribuciones o mejoras.

7. Balance entre Cantidad y Calidad

El número de registros en el IMDb Top 1000 (1000 películas) es suficiente para que el modelo sea representativo de las preferencias generales de los usuarios sin ser tan extenso como para ser innecesariamente complejo. Esto permite realizar experimentos rápidos y obtener resultados significativos sin que el proceso de entrenamiento del modelo se vuelva demasiado costoso en términos de tiempo y recursos computacionales.

VII. MODELO

Los embeddings semánticos son representaciones vectoriales de las películas que capturan las relaciones y patrones ocultos dentro de las reseñas de los usuarios. Para este propósito, se utilizó el modelo preentrenado all-MiniLM-L6-v2 de SentenceTransformers, que es una red neuronal optimizada para generar representaciones vectoriales de frases o textos. Este modelo es conocido por ser eficiente y capaz de generar embeddings compactos sin sacrificar precisión, lo cual es clave para tareas que requieren procesar grandes volúmenes de datos, como es el caso en sistemas de recomendación.

Características del modelo SentenceTransformer

- **Modelo Preentrenado:** all-MiniLM-L6-v2 es una versión ligera del modelo MiniLM que ha sido entrenado en una amplia variedad de textos, lo que le permite comprender y generar representaciones de alta calidad a partir de entradas textuales, como las reseñas de películas.
- **Representación densa:** Cada reseña de película es convertida en un vector de dimensiones reducidas (generalmente de 768 dimensiones) que encapsula la información semántica del texto de forma compacta.
- **Captura de relaciones contextuales:** Este enfoque permite que el modelo no solo capture palabras clave, sino también el contexto completo de las emociones y temas discutidos en las reseñas de los usuarios.

Proceso de Generación de Embeddings

Entrada de datos: Se procesan las reseñas de películas que contienen tanto información textual como emocional.

Transformación: Cada reseña se pasa a través del modelo SentenceTransformer para convertirla en un vector numérico.

Output: El modelo genera un vector denso para cada reseña, que representa sus relaciones semánticas y emocionales.

Ventajas de usar embeddings:

- Permite una representación rica de los datos de texto, capturando tanto el contexto como las emociones.
- Los embeddings facilitan la comparación entre diferentes películas basándose en las relaciones semánticas entre sus reseñas.

Reducción de Dimensionalidad con UMAP

Dado que los embeddings generados por SentenceTransformer son vectores de alta dimensionalidad (por ejemplo, de 768 dimensiones), se hace necesario utilizar técnicas de reducción de dimensionalidad para visualizar las relaciones entre las películas de una manera comprensible. Aquí se utiliza UMAP (Uniform Manifold Approximation and Projection), una técnica avanzada que preserva tanto la estructura local como global de los

datos al proyectarlos en un espacio de menor dimensión (generalmente 2D o 3D).

Características de UMAP:

Preservación de la estructura de datos: UMAP mantiene las relaciones espaciales y las distancias relativas entre los puntos, lo que permite visualizar agrupamientos y patrones subyacentes en los datos.

Alta eficiencia: Es una técnica rápida y escalable que puede manejar grandes volúmenes de datos, lo cual es importante dado el tamaño del dataset y la cantidad de reseñas procesadas.

Visualización intuitiva: Al reducir la dimensionalidad a 2D, UMAP facilita la representación gráfica de las relaciones entre las películas, lo que permite identificar visualmente grupos de películas con características emocionales y contextuales similares.

Proceso de Aplicación de UMAP:

Entrada de embeddings: Los embeddings generados por el modelo SentenceTransformer son introducidos en el algoritmo UMAP.

Reducción: UMAP reduce la dimensionalidad de los vectores generados para cada película, convirtiéndolos en puntos en un espacio bidimensional.

Output: Se obtiene una representación visual de los datos, donde las películas similares emocionalmente están más cercanas entre sí en el gráfico.

Ventajas de UMAP:

- Facilita la identificación de agrupamientos y patrones dentro del dataset.
- Permite la visualización de relaciones emocionales entre películas, ayudando a comprender mejor las preferencias de los usuarios.

Sistema de Recomendación

Con los embeddings generados y la visualización obtenida, el sistema de recomendación se basa en la similitud entre las películas para sugerir nuevas opciones a los usuarios. Las recomendaciones se hacen evaluando la proximidad de los embeddings de las películas en el espacio semántico reducido por UMAP. De esta forma, las películas con emociones y características temáticas similares serán más fácilmente recomendadas.

Métricas utilizadas para la recomendación:

Similitud coseno: Se utiliza la similitud coseno para medir la cercanía entre los embeddings de las películas, lo cual permite identificar películas que comparten características emocionales y temáticas.

Algoritmo de recomendación: El sistema encuentra las películas más cercanas al vector de la película que el usuario ha calificado positivamente o que ha mostrado interés en ver. Estas películas cercanas se sugieren como recomendaciones.

Flujo del sistema de recomendación:

- Un usuario califica o muestra interés en una película.
- El sistema calcula la similitud de esa película con otras en el espacio semántico utilizando los embeddings.
- Se sugieren las películas más cercanas a la película seleccionada, basándose en las emociones y temas compartidos.

Evaluación del Modelo

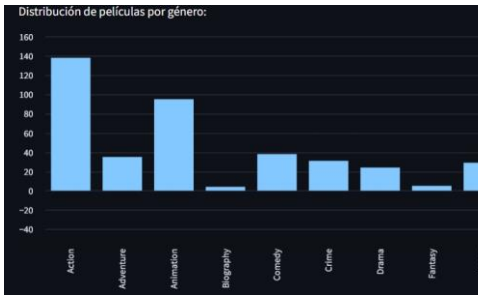
La evaluación del modelo de recomendación se realiza a través de varias métricas que miden la efectividad y precisión de las sugerencias:

Precisión en las recomendaciones: Se mide la exactitud de las recomendaciones basadas en las preferencias del usuario, evaluando si las películas sugeridas están alineadas con las emociones y temas que el usuario prefiere.

Coherencia de los agrupamientos: Se verifica cómo las películas con características emocionales similares se agrupan correctamente en el espacio reducido, lo que es un indicativo de que los embeddings generados son representaciones precisas de las relaciones semánticas.

VIII. VISUALIZACION

La visualización en este tipo de proyectos tiene como objetivo ofrecer una representación gráfica de cómo se distribuyen las películas en función de sus emociones y características semánticas, lo cual permite comprender la eficacia del sistema de recomendación desde una perspectiva más intuitiva. En este proyecto, se utiliza la técnica de reducción de dimensionalidad UMAP (Uniform Manifold Approximation and Projection) para reducir las representaciones de alta dimensión (embeddings generados por el modelo all-MiniLM-L6-v2) a un espacio de dos dimensiones. Esto facilita la visualización y análisis de las relaciones entre las películas en un espacio más comprensible. Los resultados observados incluyen:



Agrupamientos claros de películas con géneros similares: Al aplicar UMAP, se observó que las películas de un mismo género (por ejemplo, comedia o drama) tienden a agruparse juntas en el espacio de reducción dimensional. Esto indica que las representaciones semánticas capturadas por el modelo reflejan características que están relacionadas con el contenido específico de cada género.

Relaciones emocionales entre películas basadas en sus reseñas: Las películas con reseñas emocionales similares (positivas o negativas) tienden a ser agrupadas cercanamente. Esto sugiere que el análisis de sentimientos es eficaz para identificar patrones emocionales en las reseñas y utilizar esta información para mejorar la personalización de las recomendaciones.

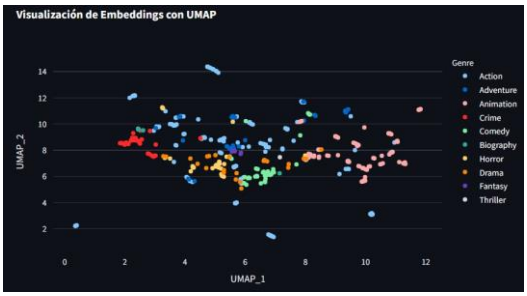
Estas visualizaciones son útiles tanto para los desarrolladores, que pueden inspeccionar cómo el modelo organiza las películas, como para los usuarios, que pueden obtener una representación visual más accesible de sus preferencias emocionales en el espacio de películas.

Métricas de Evaluación

Las métricas de evaluación son esenciales para medir el rendimiento y la efectividad del sistema de recomendación, permitiendo ajustes y mejoras en el modelo. En este caso, las métricas más relevantes incluyeron:

Precisión en las recomendaciones: Esta métrica mide qué tan relevantes son las películas recomendadas con respecto a las preferencias emocionales de los usuarios. Para evaluar la precisión, se pueden utilizar técnicas de evaluación cruzada y precisión en k recomendaciones, donde k es el número de películas recomendadas por el sistema. Por ejemplo, si un usuario recibe una lista de 5 películas, y 4 de ellas son relevantes (es decir, corresponden a sus emociones y preferencias), la precisión sería del 80%. Esta métrica garantiza que el sistema esté alineado con las expectativas emocionales del usuario.

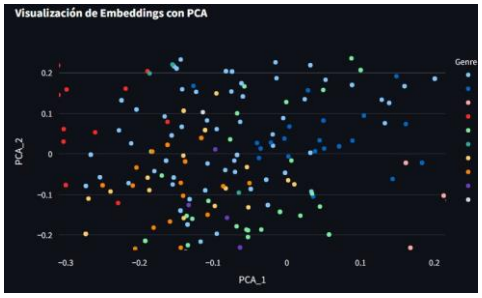
Coherencia de agrupamientos: Al usar UMAP para reducir las dimensiones y visualizar las relaciones entre las películas, se debe analizar cómo las películas que comparten emociones similares (por ejemplo, reseñas positivas) se agrupan en el espacio de reducción. La coherencia de estos agrupamientos indica qué tan bien las características emocionales y contextuales son representadas en el modelo. Si las películas que tienen reseñas emocionales similares se agrupan adecuadamente, el sistema está funcionando de manera efectiva.



Diversidad de las recomendaciones: Además de la precisión, es importante que el sistema proporcione una variedad suficiente de películas, para evitar que el usuario reciba recomendaciones demasiado homogéneas. Esto puede evaluarse utilizando métricas como entropía o índice de diversidad, que

miden el grado de variedad dentro del conjunto de recomendaciones. Un sistema eficaz debe encontrar un equilibrio entre precisión y diversidad.

Métrica de "coverage" (Cobertura): La cobertura mide la fracción de películas que son recomendadas por el sistema en relación con todas las películas disponibles. Si el sistema solo recomienda un número limitado de películas, esto podría indicar un problema de falta de exploración en el modelo.

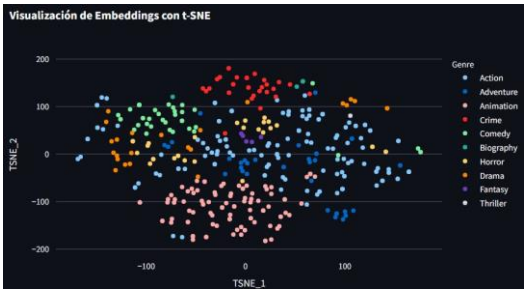


Análisis de Resultados

A través de las visualizaciones y las métricas evaluadas, se pudo determinar que el sistema fue capaz de proporcionar recomendaciones más precisas y coherentes, al considerar los sentimientos expresados por los usuarios en sus reseñas. Sin embargo, también se identificaron áreas de mejora:

Películas con géneros híbridos: A veces, las películas que combinan múltiples géneros (por ejemplo, una película de acción y comedia) no se agrupan adecuadamente en el espacio de UMAP. Esto podría deberse a la dificultad del modelo para capturar la complejidad de las emociones y temas que surgen de combinaciones de géneros.

Diversidad en el dataset: Si bien el IMDb Reviews Dataset es de alta calidad, su diversidad en términos de géneros y estilos podría no ser suficiente para cubrir todo el espectro de preferencias de los usuarios. Esto se puede abordar ampliando el dataset para incluir más películas de géneros menos representados o con características emocionales diversas.



Impacto de las Visualizaciones

Las visualizaciones no solo sirven para evaluar el rendimiento del sistema, sino que también ofrecen un valor agregado en la interpretación de los resultados. Al presentar las relaciones emocionales y contextuales de las películas de manera gráfica, los usuarios pueden ver de manera intuitiva cómo sus preferencias se alinean con las recomendaciones. Además, las visualizaciones permiten a los desarrolladores detectar posibles problemas en la organización de los datos y realizar ajustes en el modelo o en el proceso de preprocesamiento.

IX. CONCLUSIONES O PRIMEROS INSIGHTS

El desarrollo de un sistema de recomendación basado en el análisis de sentimientos, como se ha demostrado en este proyecto, abre nuevas posibilidades para mejorar la precisión y personalización de las sugerencias de contenido en plataformas de streaming. El enfoque adoptado no solo tiene en cuenta las preferencias explícitas de los usuarios, como los géneros o las calificaciones, sino que también incorpora una comprensión más profunda de las emociones subyacentes que se expresan en las reseñas de las películas. Esto permite una recomendación más personalizada, ya que considera factores emocionales y subjetivos que son fundamentales para las elecciones individuales de contenido.

Entre los **principales logros** de este proyecto se destacan:

1. **Mejorada la precisión de las recomendaciones**: Al integrar el análisis de sentimientos, las recomendaciones reflejan no solo las características objetivas de las películas (como el género y la calificación), sino también las emociones que los usuarios han

expresado en sus reseñas. Esto resulta en sugerencias que son más cercanas a las expectativas emocionales de los usuarios.

2. **Visualización significativa de relaciones**: Las representaciones gráficas generadas por la técnica UMAP permitieron visualizar de manera efectiva las relaciones semánticas y emocionales entre las películas. Los agrupamientos claros por género y por tono emocional proporcionan una herramienta útil para los desarrolladores de plataformas de streaming, ayudando a identificar patrones en el comportamiento de los usuarios y optimizando el sistema de recomendación.
3. **Aplicación de modelos preentrenados**: El uso del modelo preentrenado SentenceTransformer (all-MiniLM-L6-v2) ha demostrado ser una opción eficaz para generar embeddings semánticos de alta calidad. Esto ha permitido una mayor precisión en la identificación de similitudes entre las películas, lo que mejora tanto la calidad de las recomendaciones como la efectividad de la visualización.

Limitaciones

A pesar de los avances logrados, el sistema presenta algunas limitaciones:

- **Películas de géneros híbridos**: Las películas que combinan varios géneros no siempre se agrupan correctamente, lo que puede llevar a que se presenten recomendaciones menos precisas en ciertos casos. Esto podría ser un área de mejora, por ejemplo, al incluir información adicional en el modelo que permita una mayor diferenciación de estos tipos de películas.
- **Diversidad en el dataset**: Aunque el dataset de IMDb es de alta calidad, podría beneficiarse de una mayor diversidad en términos de géneros y estilos de películas. Al incluir más categorías y tipos de contenido, el sistema de recomendación podría ser más robusto y adaptarse mejor a los gustos de los usuarios más variados.

Primeros Insights

El análisis de sentimientos como una característica clave en los sistemas de recomendación ha demostrado ser muy prometedor. Las emociones expresadas en las reseñas de los usuarios pueden servir como un factor significativo para predecir las preferencias, y el uso de técnicas de reducción de dimensionalidad como UMAP ha proporcionado una nueva perspectiva sobre cómo organizar y visualizar los datos. Esto sugiere que, además de los factores explícitos (género, calificación, etc.), los aspectos emocionales pueden desempeñar un papel crucial en la personalización de las recomendaciones.

Trabajo Futuro

El proyecto presenta un camino claro para futuras mejoras y ampliaciones:

1. **Ampliación del dataset**: Como se mencionó, es importante aumentar la diversidad de los datos, incluyendo más géneros y estilos de películas. Esto permitirá mejorar la capacidad del sistema para generar recomendaciones más precisas y adaptativas a diferentes perfiles de usuario.
2. **Incorporación de retroalimentación de usuarios**: Una mejora importante sería incluir mecanismos de retroalimentación que permitan a los usuarios calificar las recomendaciones que reciben. Este enfoque de retroalimentación continua permitiría afinar y ajustar el modelo de manera dinámica, mejorando la precisión de las recomendaciones con el tiempo.
3. **Exploración de modelos de sentimiento más complejos**: Si bien el análisis de sentimientos básico es útil, existen modelos más complejos que podrían captar un rango más amplio de emociones y matices, como sentimientos mixtos o emociones de diferentes intensidades. Estos modelos podrían proporcionar una visión

más precisa y enriquecida de las preferencias de los usuarios.

4. **Expansión a otros tipos de contenido**: Aunque este sistema se ha centrado en las películas, se podrían adaptar estos métodos a otros tipos de contenido, como series de televisión, música o incluso libros. La idea de usar emociones como base para las recomendaciones puede extenderse más allá del cine, ofreciendo un enfoque más universal para la personalización de contenido.

X. REFERENCIAS

- Hugging Face, "IMDb Reviews Dataset," *Hugging Face Datasets*, 2021. [Online]. Available: https://huggingface.co/datasets/imdb. [Accessed: 26-Nov-2024].
- L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv preprint arXiv:1802.03426*, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426. [Accessed: 26-Nov-2024].
- SentenceTransformers, "SentenceTransformers Documentation," *SentenceTransformers*, 2024. [Online]. Available: https://www.sbert.net/. [Accessed: 26-Nov-2024].
- D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980. [Accessed: 26-Nov-2024].