الجامعة السورية الخاصة

كلية الهندسة

قسم الذكاء الصنعي وعلوم البيانات

أعدّت هذه الاطروحة

لإنجاز مقرر المشروع الفصلي في اختصاص الذكاء الصنعي وعلوم المعطيات

E-commers platform with recommendation system and semantic search

اعداد الطالبين:

حسن الحمصي

باسل ابو خبصة

أسماء المشرفين:

الدكتورة ميساء ابوقاسم

المهندسة وسام السحلي

المصلحات

المصطلح التقني	الترجمة العربية	الإختصار	المعنى
Automated Essay Scoring	التقييم الآلي للمقالات	AES	نظام يعتمد على الذكاء الصنعي لتقييم المقالات بشكل آلي
Prompt	الموجّه	-	سؤال أو توجيه يتم إعطاؤه للطلاب لكتابة مقال أو إجابة.
Prompt engineering	هندسة الموجّهات	-	تعتمد هذه النماذج على تقنيات التعلم العميق، بما في ذلك التعلم تحت الإشراف والتعلم المعزز. يتم تدريب النموذج في البداية على مجموعة بيانات كبيرة، ثم يتم تحسينه باستخدام التعليقات البشرية.
Reinforcement Learning	التعلم المعزز	RL	نوع من تعلم الآلة يتعلم السياسة المثلى عن طريق المكافآت والعقوبات، لأجل تحسين إجمالي المكافأة.
Proximal Policy Optimization	سياسة القرب	PPO	خوارزمية قوية في مجال التعلم المعزز، وتوفر استقرارًا وفعالية في تحسين سياسات النماذج.
Chain-of-Thought	تسلسل الأفكار	COT	أسلوب يستخدم تسلسل الأفكار لتوجيه عمليات التفكير وتحسين النتائج في النماذج اللغوية.
Rubrics	قواعد التقييم	-	معايير أو إرشادات تستخدم لتقييم جودة المقالات.
Trait-Based Scoring	التقييم القائم على السّمات	-	تقييم المقالات بناءً على سمات محددة مثل التنظيم والمفردات.
Rationale-based Multiple Trait Scoring	التقييم متعدد السمات القائم على الأُسس	RMTS	نهج لتقييم المقالات يعتمد على الأسس العقلية التي تبرر تصنيف القرار.
Artificial Intelligence	الذكاء الصنعي	AI	مجال من مجالات علوم الحاسب يركز على بناء أنظمة قادرة على أداء مهام تتطلب عادة ذكاءً بشرياً.

فرع من فروع الذكاء الصنعي يهتم بفهم أو توليد اللغة البشرية سواء كانت على شكل نص أو كلام.	NLP	معالجة اللغات الطبيعية	Natural language processing
مجال فرعي من تعلم الآلة يستخدم عدة طبقات مخفية في الشبكات العصبونية لحل المشكلات المعقدة عن طريق تحديد أهم السّمات للمعطيات.	DL	التعلم العميق	Deep Learning
أحد فروع الذكاء الصنعي التي تهتم بتصميم وتطوير خوارزميات وتقنيات تسمح للحواسيب بامتلاك خاصية التعلم.	ML	التعلم الآلي	Machine Learning
خوارزمية تُستخدم للتنبؤ بالقيم العددية.	SVR	انحدار المتّجه الداعم	Support Vector Regression
خوارزمية لتصنيف المعطيات.	SVM	آلة المتّجه الداعم	Support Vector Machine
خوارزمية تُستخدم لاتخاذ القرارات بناءً على الشروط والقواعد.	-	شجرة القرار	Decision Tree
خوارزمية إحصائية تُستخدم لتصنيف المعطيات إلى فئات.	LR	الانحدار اللوجستي	Logistic Regression
خوارزمية إحصائية تُستخدم للتنبؤ بالقيم العددية.	LR	الانحدار الخطّي	Linear regression
نموذج تعلم آلي يجمع بين عدة مصنفات لتحسين الأداء.	-	المصنف التجميعي	Meta-classifier
مصنف احتمالي يعتمد على نظرية بايز لتحليل المعطيات وتصنيفها.	Naive Bayes	مصنف بايز البسيط	NB
معمارية تعلم عميق فعالة لمعالجة النصوص والسياقات المعقدة.	-	المحوّلات	Transformers
نموذج لغوي مكون من عدد كبير من المعاملات	LLMs	النماذج اللغوية الكبيرة	Large Language Models

Bidirectional Encoder Representations From Transformers	تمثيلات الترميز الثنائية الاتجاه من المحولات	BERT	نموذج تعلم عميق لغوي يعتمد على بُنية المحولات لمعالجة اللغات الطبيعية وفهم السياق في النصوص.
Arabic BERT	تمثيلات الترميز الثنائية الاتجاه من المحولات خاصة باللغة العربية	AraBERT	نسخة عربية من نموذج BERT مصممة خصيصاً لمعالجة اللغة العربية.
CLAUDE 2	-	-	نموذج لغوي متقدم، يستخدم لتوليد النصوص وتحليل اللغة الطبيعية. يتميز بقدرته على فهم السياق وتقديم استجابات دقيقة وطبيعية،
Generative Pre-trained Transformer	مُحول توليدي للنصوص مسبق التدريب	GPT	نموذج لغوي قوي لتوليد النصوص وفهم اللغة.
Multi-Layers Perceptron	الشبكة العصبونية متعددة الطبقات	MLP	شبكة عصبونية تشتمل على أكثر من طبقة يستخدم لتحليل الأنماط والتصنيف.
Multi-Task Learning	التعلم متعدد المهام	MTL	تقنية تدريب نموذج واحد لأداء عدة مهام مرتبطة ببعضها.
Multi-Trait Specialization		MTS	
Data Augmentation	تعزيز المعطيات	-	أسلوب في تعلم الآلة لإنشاء معطيات جديدة عن طريق معالجة المعطيات الأصلية.
Random Oversampling	تعزيز المعطيات عشوائياً	RO	خوارزمية تستخدم لموازنة المعطيات غير المتوازنة الأصناف عن طريق تكرار أمثلة من صنف الأقلية.
Quadratic Weighted Kappa	كابا الموزونة التربيعية	QWK	مقياس إحصائي لتقييم درجة الإتّفاق بين اثنين من المقيمين أو لتقييم أداء نموذج التصنيف.
Mean Absolute Error	متوسط الخطأ المطلق	MAE	مقياس يستخدم لتقييم دقة النموذج عن طريق حساب متوسط الفرق

المطلق بين القيم المتوقعة والتنبؤ لجميع أمثلـة التدريب.			
متوسـط الخسـارة التربيعيـة لـكل مثـال، محسـوباً بقسـمة الخسـارة التربيعيـة علـى عـدد الأمثلـة.	MSE	متوسط الخطأ التربيعي	Mean Squared Error
مقياس الفرق بين القيم المتوقعة والقيم الفعلية.	RMSE	جذر متوسط الخطأ التربيعي	Root Mean Square Error

1-مقدمة و تعريف بالمشروع

1.1. مقدمة

يعاني الكثير من الأشخاص من صعوبات كبيرة عند شراء الملابس، حيث يضطرون إلى قضاء ساعات طويلة في البحث عن ما يرغبون فيه من منتجات، سواء في المتاجر التقليدية أو عبر الإنترنت. هذه العملية ليست مرهقة فحسب، بل قد تؤدي أيضًا إلى إهدار الوقت والجهد دون الوصول للنتيجة المرجوة، خاصة مع صعوبة التحكم في الوقت وكثرة الخيارات المتاحة. علاوة على ذلك، يجد العديد من العملاء صعوبة في التعبير بدقة عما يريدون، سواء كان ذلك عن طريق الكلمات أو الصور، مما يجعل تجربة التسوق أقل فعالية ويزيد من احتمالية شعورهم بالإحباط .

يهدف هذا المشروع إلى تقديم حل مبتكر لهذه المشكلة من خلال تطوير نظام ذكي يسهّل عملية التسوق ويوفر تجربة مريحة وسريعة للعملاء. يعتمد النظام على **شات بوت متقدم** يمكن للمستخدم من خلاله إرسال وصف كتابي لما يبحث عنه، أو صورة لقطعة ملابس معينة، أو حتى دمج النص مع الصورة لتحديد متطلباته بدقة. كما يقوم النظام بتحليل الطلب وتقديم منتجات مشابهة تتوافق مع رغبات المستخدم، مما يقلل الوقت المستغرق في البحث ويزيد من فرص العثور على المنتج المطلوب بسهولة وهو نظام **التوصية الهجين**,
بالاضافة لذلك يمكن للمستخدم البحث عن المنتجات باستخدام كلمات عامة تعبر عن المنتج على عكس محركات البحث التقليدية حيث كانت تعتمد على ايجاد المنتجات المطابقة للكلمة المكتوبة و ذلك باستخدام نظام **البحث الدلالي**.

هذا المشروع لا يقتصر فائدته على العملاء فقط، بل يمتد ليشمل التجار وأصحاب المتاجر، حيث يتيح لهم تقديم خدمة متميزة للعملاء، وتحسين تجربة التسوق الرقمي، وزيادة مستوى التفاعل والرضا لدى المشترين.

كما يمكن للنظام أن يسهم في تحسين كفاءة المتاجر عبر تقليل الضغط على موظفي خدمة العملاء وتسهيل إدارة الطلبات والاقتراحات المخصصة لكل مستخدم.

1.2. التعريف بالمشروع

المشروع عبارة عن نظام ذكي يعتمد على الذكاء الاصطناعي وتقنيات معالجة اللغة الطبيعية
ورؤية الحاسوب، يتيح للمستخدم البحث عن الملابس بطريقة أكثر سهولة وفعالية. يمكن للمستخدم إدخال وصف نصي، أو صورة لقطعة ملابس، أو الجمع بينهما، ليقوم النظام بتحليل
المدخلات واقتراح منتجات مشابهة تلبي احتياجاته بدقة . يُعد هذا النظام بمثابة أداة مبتكرة لتحسين تجربة التسوق عبر الإنترنت، حيث يوفر بدائل ذكية
وعملية للطرق التقليدية في البحث عن المنتجات

1.3. الأهداف

- تعزيز تجربة المستخدم عبر توفير توصيات مخصصة وبعدة أنواع بناءً على سلوك التسوق، الاهتمامات، وتفضيلات الأزياء لكل مستخدم .
- تسريع البحث عن المنتجات من خلال بوت محادثة يعتمد على معالجة اللغات الطبيعية لفهم وصف المستخدم واقتراح منتجات مطابقة.
- تقديم تجربة سلسة وتفاعلية وبسيطة تجعل من عملية التسوّق أكثر كفاءة .
- زيادة معدلات التحويل والمبيعات عبر تحسين دقة عرض المنتجات الملائمة لاحتياجات العملاء.
- الاستفادة من تحليلات البيانات لتتبع أنماط الشراء وسلوك المستخدمين بهدف تطوير الاستراتيجيات التسويقية .
- دعم سهولة الاستخدام والمرونة عبر واجهات تفاعلية تدمج مختلف تسهيلات العمليات.
- استقطاب الجمهور بجميع الفئات والاهتمامات وتوفير جميع الاحتياجات بأسهل الوسائل.
- رفع مستوى التنافسية من خلال دمج الذكاء الاصطناعي مما يجعل المتجر متميزاً عن المتاجر التقليدية.
- ضمان إعجاب العميل بالمنتج المباع من المرة الأولى دون خسارة رضا الزبائن وإعادة المنتجات المباعة.

1.4. المهمّة

تتمثل مهمة المشروع في تقديم حل عملي وفعّال لتحديات التسوق الرقمي من خلال :

- تمكين العملاء من الوصول إلى المنتجات المطلوبة بسهولة وسرعة .
- تقليل الجهد والوقت الضائع في البحث التقليدي عن الملابس .
- تحسين تجربة التسوق الرقمية وزيادة رضا العملاء .
- دعم استراتيجيات التجارة الإلكترونية من خلال نظام ذكي وفعال,

1.5. رؤية المشروع

تسعى رؤية المشروع إلى إحداث تحول نوعي في تجربة التسوق عبر الإنترنت من خلال منصة
ذكية متكاملة. تهدف الرؤية إلى :

- جعل عملية البحث عن الملابس أكثر دقة ومرونة .
- تعزيز مكانة النظام كأداة أساسية للتسوق الرقمي المبتكر .
- المساهمة في تطوير مستقبل التجارة الإلكترونية بالاعتماد على تقنيات الذكاءالاصطناعي

2- الدراسة المرجعية

شهدت منصّات التجارة الإلكترونية خلال السنوات الأخيرة تطورًا ملحوظًا في توظيف تقنيات الذكاء الاصطناعي بهدف تحسين تجربة المستخدم وزيادة فعالية الوصول إلى المنتجات المناسبة. ويُعد كلٌّ من أنظمة التوصية (Recommendation Systems) والبحث الدلالي (Semantic Search) من أكثر المكوّنات تأثيرًا في نجاح هذه المنصّات، لما لهما من دور محوري في فهم سلوك المستخدم ونيّته وتقديم محتوى ومنتجات ملائمة بشكل شخصي ودقيق.

أظهرت الدراسات الحديثة أن الاعتماد على الأساليب التقليدية، مثل الترشيح التعاوني أو البحث القائم على الكلمات المفتاحية فقط، لم يعد كافيًا للتعامل مع التحديات المعقّدة التي تفرضها البيئات التجارية الرقمية، كندرة البيانات (Data Sparsity)، ومشكلة المستخدم أو المنتج الجديد (Cold Start)، وضعف فهم المعنى الدلالي لاستفسارات المستخدم. ومن هنا برزت الأنظمة الهجينة (Hybrid Systems) التي تجمع بين أكثر من تقنية، مثل دمج الترشيح التعاوني مع الترشيح القائم على المحتوى أو تحليل المشاعر، إضافةً إلى الاستفادة من النماذج العميقة مثل الشبكات العصبية والمشفّرات التلقائية (Autoencoders)، لما أظهرته من قدرة عالية على تحسين دقة التوصيات وجودتها في سياقات التجارة الإلكترونية

.

بالتوازي مع ذلك، شهد مجال البحث الدلالي تطورًا كبيرًا بفضل استخدام النماذج اللغوية العميقة والمحولات (Transformers)، مثل BERT وGPT، والتي مكّنت أنظمة البحث من الانتقال من المطابقة اللفظية إلى فهم المعنى والسياق. وقد أثبتت الأبحاث أن دمج التمثيلات الدلالية الكثيفة (Dense Embeddings) مع محركات بحث متجهية ضمن نماذج هجينة يحقق تحسنًا ملحوظًا في مقاييس الترتيب والاسترجاع، مثل nDCG وMRR، مقارنةً بالأنظمة التقليدية المعتمدة على BM25 فقط، خاصةً في البحث عن المنتجات داخل منصّات التجارة الإلكترونية واسعة النطاق .

كما تشير الدراسات الحديثة إلى أن الدمج بين أنظمة التوصية والبحث الدلالي يوفّر تجربة أكثر تكاملًا للمستخدم، حيث لا يقتصر النظام على اقتراح منتجات بناءً على السلوك السابق فحسب، بل يصبح قادرًا على فهم الاستفسارات النصية المعقّدة، وتحليل تفضيلات المستخدم الضمنية، وتقديم نتائج ذات صلة أعلى من الناحية الدلالية والسلوكية. هذا التكامل يُعد توجّهًا بحثيًا معاصرًا، تدعمه نتائج تطبيقات صناعية حقيقية في شركات كبرى، مثل Amazon وWalmart، والتي أثبتت أثره الإيجابي على معدلات النقر والتحويل ورضا المستخدم.

بناءً على ما سبق، يهدف هذا الفصل إلى استعراض وتحليل مجموعة من الدراسات والأبحاث المرجعية التي تناولت أنظمة التوصية الهجينة وتقنيات البحث الدلالي في سياق التجارة الإلكترونية، مع التركيز على النماذج المستخدمة، والبيانات المعتمدة، ونتائج التقييم. ويشكّل

هذا الاستعراض الأساس العلمي الذي يُبنى عليه تصميم وتنفيذ المنصّة المقترحة في هذا المشروع، بما يضمن توافقها مع أحدث الاتجاهات البحثية والتطبيقية في هذا المجال.

1.2 الدراسة المرجعية بالنسبة لأنظمة التوصية الهجينة (Hybrid Recommendation Systems)

بدأت أنظمة التوصية (Recommender Systems) في مراحلها الأولى بالاعتماد على أساليب منفصلة، أبرزها **الترشيح التعاوني (Collaborative Filtering)** الذي يستند إلى تفاعلات المستخدمين مع العناصر، و**الترشيح القائم على المحتوى (Content-Based Filtering)** الذي يعتمد على خصائص العناصر نفسها. ورغم النجاح الأولي لهذه الأساليب، إلا أنها واجهت تحديات جوهرية حدّت من فعاليتها في البيئات الواقعية، مثل مشكلة ندرة البيانات (Data Sparsity)، ومشكلة المستخدم أو العنصر الجديد (Cold Start)، وضعف القدرة على تمثيل التفضيلات المعقدة والمتغيرة للمستخدمين، خاصة في منصّات التجارة الإلكترونية واسعة النطاق.

استجابةً لهذه التحديات، اتجهت الأبحاث إلى تطوير **أنظمة التوصية الهجينة** التي تهدف إلى دمج مزايا أكثر من نهج توصية واحد ضمن إطار موحّد، بغية تحسين دقة التوصيات وزيادة تغطيتها واستقرارها. في المراحل المبكرة، تم الاعتماد على أساليب هجينة بسيطة تعتمد على الدمج الخطي أو الترجيح بين مخرجات الترشيح التعاوني والترشيح القائم على المحتوى، أو على التبديل الشرطي بينهما تبعًا لتوافر البيانات. ومع ذلك، ظلّ هذا النوع من الدمج محدود القدرة على نمذجة العلاقات غير الخطية بين المستخدمين والعناصر.

مع تطور التعلم العميق، شهدت أنظمة التوصية الهجينة قفزة نوعية من خلال توظيف **الشبكات العصبية العميقة (DNNs)** و**المشفّرات التلقائية (Autoencoders)**، التي مكّنت من تعلم تمثيلات كامنة (Latent Representations) غنية وقادرة على دمج مصادر متعددة من البيانات، مثل التفاعلات التاريخية، وخصائص المستخدم، وسمات المنتجات. أظهرت دراسات مبكرة أن استخدام Autoencoders ضمن أطر هجينة ساهم في تحسين أداء التوصية في البيئات ذات البيانات المتناثرة، كما عزّز قدرة النظام على التعميم والتعامل مع المستخدمين الجدد.

لاحقًا، ركّزت الأبحاث الحديثة على تطوير نماذج هجينة أكثر تعقيدًا تجمع بين **التعلّم التمثيلي العميق** و**الإشارات السلوكية والاجتماعية**. فقد اقترحت بعض الدراسات نماذج هجينة قائمة على الشبكات العصبية العميقة المدعومة بشبكات علاقات المستخدم–المستخدم (User-User Networks)، حيث يتم دمج العلاقات الاجتماعية أو التشابه السلوكي بين المستخدمين مع التفاعلات التقليدية مع العناصر، مما أظهر تحسنًا ملحوظًا في مقاييس الخطأ مثل RMSE وMAE، خاصة في سيناريوهات Cold Start. كما أسهمت نماذج

Variational Autoencoders (VAE) المطوّرة، مثل Variational Bandwidth Auto-Encoder، في تحسين تمثيل التفضيلات الكامنة من خلال آليات تكيفية تزيد من مرونة النموذج وقدرته على التعميم.

إلى جانب ذلك، توسّعت أنظمة التوصية الهجينة لتشمل **المعلومات الدلالية المستخرجة من النصوص**، لا سيما مراجعات المستخدمين (Product Reviews). فقد أظهرت دراسات حديثة أن دمج تحليل المشاعر (Sentiment Analysis) باستخدام نماذج لغوية عميقة مثل BiLSTM أو BERT ضمن الأطر الهجينة يضيف بُعدًا دلاليًا يعكس الرأي الحقيقي للمستخدم، وليس فقط تقييمه العددي، مما أدى إلى تحسين ملحوظ في مقاييس Precision@K وRecall@K وF1-score، خصوصًا في منصّات التجارة الإلكترونية الغنية بالمراجعات النصية.

لقياس أداء أنظمة التوصية الهجينة، تعتمد الأبحاث على مجموعات بيانات معيارية شهيرة مثل Amazon Product Reviews، MovieLens، وLast.fm، إضافة إلى مجموعات بيانات خاصة ببيئات التجارة الإلكترونية. وتُستخدم مقاييس تقييم متنوعة، من بينها RMSE وMAE لقياس دقة التنبؤ بالتقييمات، ومقاييس ترتيبية مثل NDCG@K وRecall@K وPrecision@K لتقييم جودة التوصيات من منظور تجربة المستخدم. وقد أظهرت النتائج في معظم الدراسات تفوق الأنظمة الهجينة على النماذج الأحادية، خاصة في الحالات التي تعاني من نقص البيانات أو تباين سلوك المستخدمين.

ورغم هذا التقدم، لا تزال أنظمة التوصية الهجينة تواجه تحديات بحثية مفتوحة، من أبرزها **قابلية التفسير**، إذ تُعد النماذج العميقة المستخدمة بمثابة "صناديق سوداء" يصعب تفسير أسباب توصياتها، إضافةً إلى التعقيد الحسابي وارتفاع كلفة التدريب، وصعوبة الموازنة بين مصادر البيانات المختلفة دون إدخال تحيزات غير مرغوبة. ولمعالجة هذه القضايا، اتجهت الأبحاث الحديثة إلى توظيف استراتيجيات مثل **التعلم متعدد المهام (Multi-Task Learning)**، ودمج آليات الانتباه (Attention Mechanisms) لتحديد الإشارات الأكثر تأثيرًا في التوصية، إضافةً إلى تطوير نماذج هجينة أكثر تكاملًا مع أنظمة البحث الدلالي.

تشير الاتجاهات البحثية الحديثة إلى أن المستقبل القريب لأنظمة التوصية الهجينة يتجه نحو **الدمج العميق بين التوصية والبحث الدلالي** ضمن منصّات التجارة الإلكترونية، بحيث لا يقتصر النظام على اقتراح المنتجات بناءً على السلوك السابق فحسب، بل يصبح قادرًا على فهم نية المستخدم الدلالية واستفساراته النصية، وتقديم توصيات وسياقات بحث متكاملة تعزز تجربة المستخدم وترفع معدلات التفاعل والتحويل. ويشكّل هذا التوجّه الأساس العلمي الذي يستند إليه هذا المشروع في تصميم منصّة تجارة إلكترونية تجمع بين نظام توصية هجين ومحرك بحث دلالي متقدم.

Paper #	Publication Date	Models/Techniques	Dataset	Evaluation Metrics	Results
1	2023	Deep Neural Network + User-User Graph	Benchmark CF datasets	RMSE, MAE, F1	+19% RMSE, +9.2% MAE
2	2022	Variational Bandwidth Autoencoder (VBAE)	Movies, Music, E-commerce	NDCG@10, Recall@10	+8.7% NDCG@10, +6.4% Recall@10
3	2022–2023	Sentiment Model (BiLSTM/BERT) + MF Hybrid	Amazon Reviews	Precision@10, Recall@10, F1	+7.8% P@10, +6.2% R@10
4	2016	Autoencoders + MF/CF	MovieLens, Last.fm	RMSE, Precision@K	+5–8% RMSE, +6% ranking
5	2019	Hybrid CF + Content-based + Clustering	Library system data	Precision, Recall, Hit Rate	Improved accuracy & coverage

2.2 الدراسة المرجعية بالنسبة للبحث الدلالي (Semantic Search)

بدأت أنظمة البحث في مراحلها الأولى بالاعتماد على الأساليب التقليدية القائمة على **المطابقة اللفظية (Lexical Matching)**، مثل نماذج استرجاع المعلومات الكلاسيكية TF-IDF ولاحقًا BM25، والتي تعتمد على تكرار الكلمات وتوزيعها داخل الوثائق لتحديد درجة الصلة بين الاستعلام والنتائج. ورغم بساطة هذه النماذج وكفاءتها الحسابية، إلا أنها تعاني من قصور جوهري يتمثل في عدم قدرتها على فهم **المعنى الدلالي** للاستعلامات، حيث تفشل في التعامل مع المرادفات، والسياق، وتنوّع الصياغات اللغوية، مما يؤدي إلى نتائج غير دقيقة، خاصة في بيئات التجارة الإلكترونية التي تتسم بتنوع أوصاف المنتجات واختلاف نية المستخدم.

مع تطور تقنيات معالجة اللغة الطبيعية (NLP)، بدأت الأبحاث بالانتقال نحو **البحث الدلالي**، الذي يهدف إلى تمثيل كل من الاستعلامات والوثائق ضمن فضاء دلالي مشترك يعكس المعنى وليس فقط الشكل اللفظي. في المراحل الأولى، تم استخدام نماذج التضمين الكلمي مثل Word2Vec وGloVe لتمثيل الكلمات، ثم تجميعها لتمثيل الجمل أو الوثائق. ورغم أن هذه الأساليب حسّنت من القدرة على التقاط العلاقات الدلالية البسيطة، إلا أنها كانت محدودة بسبب تمثيلها الثابت للكلمات، وعدم قدرتها على مراعاة السياق الذي تظهر فيه الكلمة داخل الجملة.

شكّل ظهور **نماذج المحوّلات (Transformers)**، مثل BERT وRoBERTa وSentence-BERT، نقطة تحوّل رئيسية في مجال البحث الدلالي، حيث تعتمد هذه النماذج على آلية **الانتباه الذاتي (Self-Attention)** لتعلّم تمثيلات سياقية ديناميكية على مستوى الجملة أو النص الكامل. سمحت هذه التمثيلات بتحويل كل من الاستعلامات والوثائق إلى **تضمينات كثيفة (Dense Embeddings)** يمكن مقارنتها باستخدام مقاييس تشابه متقدمة، مثل Cosine Similarity، مما أدى إلى تحسين ملحوظ في جودة الاسترجاع مقارنة بالبحث القائم على الكلمات المفتاحية فقط.

اعتمدت الأبحاث الحديثة على نماذج **Dense Retrieval**، حيث يتم فصل عملية الاسترجاع إلى مرحلتين: الأولى تعتمد على استرجاع أوّلي سريع باستخدام التضمينات الدلالية، والثانية تستخدم نماذج أكثر تعقيدًا مثل Cross-Encoders لإعادة ترتيب النتائج بدقة أعلى. ومع ذلك، أظهرت الدراسات أن نماذج الاسترجاع الكثيف قد تفشل أحيانًا في التقاط المطابقات اللفظية الدقيقة، خاصة عند البحث عن أسماء منتجات أو رموز محددة.

ولمعالجة هذه الإشكالية، ظهر اتجاه **النماذج الهجينة (Hybrid Retrieval)** التي تدمج بين الإشارات الدلالية الكثيفة والإشارات اللفظية التقليدية مثل BM25، مما أدى إلى تحسين متوازن في كل من الدقة والاستدعاء.

في سياق التجارة الإلكترونية، أثبت البحث الدلالي فعاليته الكبيرة في **فهم نية المستخدم (User Intent)** والتعامل مع الاستعلامات الغامضة أو الطويلة، مثل البحث الوصفي عن المنتجات. وقد أظهرت دراسات تطبيقية واسعة النطاق أن استخدام التضمينات الدلالية المستخرجة من نماذج لغوية كبيرة (LLMs) في البحث عن المنتجات يؤدي إلى تحسين ملحوظ في مقاييس الترتيب مثل nDCG@10 وMRR، إضافة إلى تحسين مؤشرات الأعمال الفعلية مثل معدل النقر (CTR) ومعدل التحويل (Conversion Rate). كما تم

توظيف محركات بحث متجهية مثل FAISS وWeaviate وElasticsearch Vector Search لدعم الاسترجاع الدلالي على نطاق واسع.
إلى جانب البحث النصي، توسّعت الأبحاث الحديثة نحو **البحث الدلالي متعدد الوسائط (Multimodal Semantic Search)**، الذي يدمج بين النصوص والصور، خاصة في مجالات مثل الأزياء والتجارة البصرية. تعتمد هذه النماذج على معماريات ثنائية المرمّز (Dual-Encoder) مثل CLIP، التي توحّد تمثيل النص والصورة في فضاء دلالي مشترك، مما يمكّن المستخدم من البحث عن المنتجات باستخدام أوصاف نصية أو صور مرجعية، وأثبتت هذه المقاربة تفوقها على البحث النصي الأحادي في مقاييس الدقة.
لقياس أداء أنظمة البحث الدلالي، تعتمد الدراسات على مجموعات بيانات معيارية مثل MS MARCO وBEIR وTREC Deep Learning، إضافة إلى مجموعات بيانات خاصة بالتجارة الإلكترونية. وتُستخدم مقاييس تقييم ترتيبية مثل nDCG@K وMRR وRecall@K وPrecision@K، والتي تعكس جودة ترتيب النتائج ومدى توافقها مع نية المستخدم. وقد أظهرت النتائج تفوقًا واضحًا للأنظمة الدلالية والهجينة على الأنظمة اللفظية التقليدية، خاصة في الاستعلامات المعقدة أو غير المباشرة.
ورغم هذا التقدم، لا تزال أنظمة البحث الدلالي تواجه تحديات بحثية مهمة، من أبرزها **قابلية التفسير**، إذ يصعب توضيح سبب اختيار نتائج معينة ضمن النماذج العميقة، إضافةً إلى ارتفاع الكلفة الحسابية للتدريب والاسترجاع، والحاجة إلى التكيّف مع مجالات جديدة تعاني من نقص البيانات المشروحة. ولمعالجة هذه القضايا، اتجهت الأبحاث الحديثة إلى استخدام **التعلم الذاتي (Self-Supervised Learning)**، واستراتيجيات **توسيع الاستعلامات (Query Expansion)** باستخدام النماذج اللغوية الكبيرة، بالإضافة إلى تحسين كفاءة الاسترجاع عبر هياكل فهرسة متقدمة.
تشير الاتجاهات البحثية المستقبلية إلى أن البحث الدلالي سيتجه نحو **تكامل أعمق مع النماذج اللغوية الكبيرة (LLMs)** لتقديم تجارب بحث تفاعلية، قادرة على تفسير النتائج وتخصيصها حسب المستخدم، إلى جانب دمجه بشكل وثيق مع أنظمة التوصية داخل منصّات التجارة الإلكترونية. ويُعد هذا التكامل حجر الأساس في تصميم الأنظمة الحديثة التي تسعى إلى تقديم تجربة مستخدم ذكية، دقيقة، وشاملة، وهو ما يستند إليه هذا المشروع في بناء محرك بحث دلالي متقدم ضمن منصّة تجارة إلكترونية متكاملة.

Paper #	Publication Date	Models/Techniques	Dataset	Evaluation Metrics	Results
1	2023	LLM embeddings (GPT/BERT), Hybrid lexical-semantic retrieval	Amazon internal product data	nDCG@10, MRR	+15% nDCG@10, +13% MRR
2	2023	BERT dense retriever + BM25 hybrid (LADR)	MS MARCO, NQ, TREC DL	Recall@100, MRR	+12–18% Recall, +10% MRR
3	2024	Sentence-BERT, ColBERTv2, multilingual encoders	Walmart multilingual catalog	CTR, Conversion Rate	+8–11% CTR, +10% conversions
4	2024	DistilBERT, RoBERTa with contrastive & self-supervision	MS MARCO, domain corpora	MRR@10	+9% MRR@10
5	2025	CLIP, BERT, ViT multimodal dual-encoder	Amazon Products, FashionIQ	Precision@5	+14% Precision@5
6	2025	GPT-based query expansion + BERT retriever	MS MARCO v2, BEIR	nDCG@10, Recall	+6.2% nDCG@10, +7.8% Recall

3- مقاييس التقييم (Evaluation Metrics)

في أنظمة **التوصية والبحث الدلالي** ضمن منصّات التجارة الإلكترونية، تُعدّ عملية تقييم الأداء عنصرًا أساسيًا للحكم على جودة النموذج ومدى فعاليته في تلبية احتياجات المستخدم. ويعتمد التقييم على قياس قدرة النظام على تقديم توصيات دقيقة، مرتّبة بشكل مناسب، ومتوافقة مع تفضيلات المستخدم أو مع نية البحث الدلالية. ونظرًا لاختلاف طبيعة المهام بين **التوصية** و**الاسترجاع الدلالي**، يتم استخدام مجموعة من المقاييس الكمية التي تعكس دقة التنبؤ وجودة الترتيب وفعالية الاسترجاع.

مقياس متوسط الخطأ المطلق (MAE)

يُستخدم مقياس **MAE (Mean Absolute Error)** بشكل شائع في تقييم **أنظمة التوصية القائمة على التنبؤ بالتقييمات**، حيث يقيس متوسط الفرق المطلق بين القيم المتوقعة من النظام والقيم الفعلية التي قدّمها المستخدمون.

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m} |h(x^i) - y^i|$$

يمتاز هذا المقياس بسهولة تفسيره، إذ تعبّر قيمته مباشرة عن مقدار الخطأ المتوسط في التنبؤ. وكلما اقتربت قيمة MAE من الصفر، دلّ ذلك على قدرة أعلى للنموذج على تمثيل تفضيلات المستخدمين بدقة. ويُعد مناسبًا لقياس الأداء العام دون تضخيم أثر القيم الشاذة.

مقياس الجذر التربيعي لمتوسط مربع الخطأ (RMSE)

يُعد **RMSE (Root Mean Squared Error)** من أهم المقاييس المستخدمة في تقييم دقة التنبؤ في أنظمة التوصية، خاصة عند الرغبة في إعطاء وزن أكبر للأخطاء الكبيرة.

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} (h(x^i) - y^i)^2}$$

يتميّز RMSE بحساسيته العالية للأخطاء الكبيرة مقارنة بـ MAE، مما يجعله مناسبًا لتقييم النماذج التي يُراد منها تقليل الانحرافات الحادة في التوصيات، خصوصًا في السيناريوهات التي تؤثر فيها التوصيات الخاطئة بشكل كبير على تجربة المستخدم.

مقياس الدقة عند K (Precision@K)

يُستخدم **Precision@K** لتقييم جودة التوصيات المرتّبة، حيث يقيس نسبة العناصر ذات الصلة ضمن أعلى K عنصر موصى به للمستخدم.

$$Precision@K = \frac{\text{عدد العناصر ذات الصلة ضمن أعلى } K}{K}$$

يعكس هذا المقياس مدى قدرة النظام على تقديم توصيات دقيقة في المراتب الأولى، وهو مهم جدًا في أنظمة التجارة الإلكترونية، حيث غالبًا ما يتفاعل المستخدم مع عدد محدود من النتائج المعروضة.

مقياس الاستدعاء عند K (Recall@K)

يقيّم **Recall@K** قدرة النظام على استرجاع جميع العناصر ذات الصلة للمستخدم ضمن أعلى K نتيجة.

$$Recall@K = \frac{\text{عدد العناصر ذات الصلة ضمن أعلى } K}{\text{إجمالي العناصر ذات الصلة}}$$

يُستخدم هذا المقياس لتحديد مدى شمولية النظام في تقديم التوصيات، ويُعد مكملًا لمقياس Precision@K، حيث يوازن بين الدقة والتغطية.

مقياس جودة الترتيب التراكمي الموزون (nDCG@K)

يُعد **nDCG@K (Normalized Discounted Cumulative Gain)** من أهم المقاييس المستخدمة في كل من **أنظمة التوصية والبحث الدلالي**، إذ لا يكتفي بتحديد ما إذا كانت العناصر صحيحة أم لا، بل يأخذ بعين الاعتبار **ترتيبها داخل القائمة**.

$$DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i}-1}{log_2(i+1)}, nDCG@K = \frac{DCG@K}{IDCG@K}$$

حيث تمثل rel_i درجة الصلة للعنصر في المرتبة i، بينما يمثل $IDCG@K$ أفضل ترتيب ممكن. كلما اقتربت قيمة nDCG@K من 1، دلّ ذلك على جودة عالية في ترتيب النتائج بما يتوافق مع تفضيلات المستخدم أو نية البحث.

مقياس متوسط الرتبة العكسية (MRR)

يُستخدم **MRR (Mean Reciprocal Rank)** بشكل خاص في **أنظمة البحث الدلالي**، لقياس سرعة وصول النظام إلى أول نتيجة ذات صلة.

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q} \frac{1}{rank_i}$$

حيث تمثل $rank_i$ موضع أول نتيجة صحيحة للاستعلام i. يُعد هذا المقياس مهمًا في تقييم محركات البحث، إذ يعكس تجربة المستخدم في الوصول السريع إلى النتيجة المطلوبة.

4- المعطيات الأولية وطرق تحصيلها

1.4 المعطيات الأولية الخاصة بمنصّة التجارة الإلكترونية

- **مجموعة معطيات: (AMAZON_FASHION)**

تُعد مجموعة معطيات **AMAZON_FASHION** من أشهر مجموعات البيانات المستخدمة في أبحاث **أنظمة التوصية** في مجال التجارة الإلكترونية، وهي جزء من مجموعات Amazon Product Data التي تم تجميعها من منصّة Amazon على مدار عدة سنوات. تحتوي هذه المجموعة على تفاعلات المستخدمين مع منتجات الأزياء، وتشمل **تقييمات رقمية (Ratings)**، ومراجعات نصية (Reviews)، وبيانات زمنية تعبّر عن سلوك المستخدم الفعلي أثناء التفاعل مع المنتجات.

تتضمن مجموعة البيانات ملايين التفاعلات التي تمثل عمليات تقييم قام بها مستخدمون حقيقيون لمنتجات أزياء متنوعة، مثل الملابس، الأحذية، الإكسسوارات، والحقائب. ويحتوي كل سجل تفاعل عادة على:

- معرف المستخدم (User ID)
- معرف المنتج (Product ID / ASIN)
- التقييم الرقمي (عادة من 1 إلى 5)
- نص المراجعة (Review Text)
- الطابع الزمني (Timestamp)

يمثل هذا التنوع في التفاعلات أساسًا غنيًا لتدريب وتقييم **أنظمة التوصية الهجينة**، حيث يتيح الجمع بين:

- الإشارات التعاونية (Collaborative Signals) المستخلصة من التقييمات،
- والإشارات الدلالية (Semantic Signals) المستخلصة من المراجعات النصية.

كما أن طبيعة البيانات تعكس تحديات واقعية شائعة في التجارة الإلكترونية، مثل **ندرة البيانات**، وعدم توازن التقييمات، وتباين نشاط المستخدمين، مما يجعلها مناسبة لاختبار متانة النماذج المقترحة في سيناريوهات حقيقية.

- **مجموعة معطيات: (meta_AMAZON_FASHION)**

تمثل مجموعة **meta_AMAZON_FASHION** البيانات الوصفية (Metadata) المرتبطة بمنتجات الأزياء في Amazon، وهي مكملة لمجموعة AMAZON_FASHION، وتُستخدم بشكل أساسي في **أنظمة التوصية الهجينة** و**محركات البحث الدلالي**.

تحتوي هذه المجموعة على معلومات تفصيلية لكل منتج، من بينها:

- عنوان المنتج (Product Title)
- وصف المنتج (Product Description)

- الفئة أو الفئات (Categories)
- العلامة التجارية (Brand)
- الصور (Image URLs)
- السعر (Price)
- الخصائص الفنية أو الوصفية (Attributes)

تُعد هذه البيانات ضرورية لبناء **نماذج توصية قائمة على المحتوى (Content-Based)**، حيث تُمكّن النظام من فهم خصائص المنتج ومعناه الدلالي، بدل الاعتماد فقط على تفاعلات المستخدمين. كما تُستخدم أوصاف المنتجات والعناوين في بناء **تضمينات دلالية (Embeddings)** تُستثمر في محركات البحث الدلالي، مما يسمح بفهم استعلامات المستخدم النصية وربطها بالمنتجات الأكثر صلة.

أهمية الدمج بين مجموعتي البيانات

يُشكّل الدمج بين **AMAZON_FASHION** و**meta_AMAZON_FASHION** حجر الأساس في هذا المشروع، حيث يتيح بناء نظام متكامل يجمع بين:

- **الترشيح التعاوني** المعتمد على سلوك المستخدم،
- **الترشيح القائم على المحتوى** المعتمد على خصائص المنتج،
- **البحث الدلالي** القادر على فهم نية المستخدم النصية.

يسمح هذا التكامل بتطوير **نظام توصية هجين (Hybrid Recommendation System)** قادر على التعامل مع مشاكل المستخدم الجديد والمنتج الجديد، وتحسين دقة التوصيات في حالات ندرة البيانات. كما يمكّن من بناء **محرك بحث دلالي** يتجاوز حدود البحث بالكلمات المفتاحية، من خلال مطابقة المعنى والسياق بين استعلام المستخدم وأوصاف المنتجات.

الخصائص الإحصائية والتحديات

تتميّز هذه المجموعات بعدة خصائص تجعلها بيئة بحثية غنية:

- الحجم الكبير للبيانات، مما يتطلب تقنيات فعّالة في المعالجة والتخزين.
- عدم التوازن في توزيع التقييمات، حيث تميل التقييمات الإيجابية إلى الهيمنة.
- وجود نصوص مراجعات غير منظمة، تتفاوت في الطول والجودة.
- تنوع كبير في فئات المنتجات وأنماط المستخدمين.

تفرض هذه الخصائص تحديات بحثية مهمة، مثل الحاجة إلى تقنيات تنظيف بيانات متقدمة، واستخراج تمثيلات دلالية دقيقة، واختيار مقاييس تقييم مناسبة تعكس جودة التوصية والاسترجاع الدلالي.

دور هذه المعطيات في هذا المشروع

في هذا المشروع، تم الاعتماد على:

- **AMAZON_FASHION** لتعلّم أنماط تفضيلات المستخدمين وبناء نموذج توصية يعتمد على التفاعلات الفعلية،
- **meta_AMAZON_FASHION** لاستخراج التمثيلات الدلالية للمنتجات ودعم البحث الدلالي والتوصية القائمة على المحتوى.

يضمن هذا الاختيار توافق المنظومة المقترحة مع سيناريوهات حقيقية في التجارة الإلكترونية، ويعزز قابلية تعميم النتائج وإمكانية تطبيق النظام في بيئات إنتاجية مستقبلية.

5- الدراسة النظرية

1.5 التعلم الالي

يُعد التعلّم الآلي أحد فروع الذكاء الاصطناعي، ويهتم بتطوير خوارزميات ونماذج قادرة على التعلّم من البيانات وتحسين أدائها تلقائيًا دون الحاجة إلى برمجة صريحة لكل حالة. يعتمد التعلّم الآلي على تحليل الأنماط والعلاقات داخل البيانات بهدف التنبؤ أو اتخاذ القرار.

2.5 اهمية التعلم الالي

تكمن أهمية التعلّم الآلي في قدرته على التعامل مع كميات كبيرة من البيانات واستخلاص المعرفة منها، مما جعله يُستخدم على نطاق واسع في العديد من المجالات مثل:

- الطب (تشخيص الأمراض)
- الاقتصاد والأعمال (التنبؤ بالأسعار وسلوك العملاء)
- الأمن السيبراني
- معالجة الصور والنصوص
- أنظمة التوصية مثل (YouTube, Netflix)

3.5 انواع التعلم الالي

ينقسم التعلّم الآلي إلى عدة أنواع رئيسية، من أهمها:

1.3.5 التعلّم الخاضع للإشراف (Supervised Learning)

يعتمد على بيانات مُعلَّمة (Labeled Data)، حيث يتم تدريب النموذج باستخدام مدخلات ومخرجات معروفة مسبقًا، مثل:

- الانحدار الخطي (Linear Regression)
- أشجار القرار (Decision Trees)
- الشبكات العصبية

2.3.5 التعلّم غير الخاضع للإشراف (Unsupervised Learning)

يُستخدم مع بيانات غير مُعلَّمة، ويهدف إلى اكتشاف الأنماط والعلاقات، مثل:

- التجميع (Clustering)
- تقليل الأبعاد (Dimensionality Reduction)

3.3.5 التعلّم المعزز (Reinforcement Learning)

يعتمد على مبدأ المكافأة والعقوبة، حيث يتعلم النظام من خلال التفاعل مع البيئة، ويُستخدم بكثرة في:

- الألعاب
- الروبوتات
- الأنظمة الذكية

4.5 خطوات التعلم الالي

تمر عملية التعلّم الآلي بعدة مراحل، منها:

- جمع البيانات
- معالجة البيانات وتنظيفها
- اختيار النموذج المناسب
- تدريب النموذج
- اختبار وتقييم الأداء
- تحسين النموذج

5.5 معالجة اللغات الطبيعية

تُعد معالجة اللغات الطبيعية أحد فروع الذكاء الاصطناعي والتعلّم الآلي، وتهدف إلى تمكين الحاسوب من فهم اللغة البشرية المكتوبة أو المنطوقة وتحليلها والتفاعل معها بطريقة تحاكي الفهم البشري. تجمع تقنيات الـ NLP بين علوم الحاسوب واللغويات والإحصاء بهدف معالجة النصوص واستخلاص المعاني منها.

6.5 اهمية معالجة اللغات الطبيعية

تكتسب معالجة اللغات الطبيعية أهمية كبيرة بسبب الانتشار الواسع للبيانات النصية، مثل الرسائل، المقالات، ومنشورات وسائل التواصل الاجتماعي. ومن أبرز فوائدها:

- تحسين التفاعل بين الإنسان والحاسوب
- تحليل كميات ضخمة من النصوص بسرعة ودقة
- دعم اتخاذ القرار بناءً على المحتوى النصي
- أتمتة المهام اللغوية التي تتطلب وقتًا وجهدًا بشريًا كبيرًا

7.5 مراحل معالجة اللغات الطبيعية

- تمر عملية معالجة النصوص بعدة مراحل أساسية، منها:
- **المعالجة المسبقة للنصوص**: مثل إزالة الرموز غير المهمة، وتحويل النص إلى صيغة موحدة.
- **تجزئة النص (Tokenization)**: تقسيم النص إلى كلمات أو جمل.
- **إزالة كلمات التوقف (Stop Words)**: حذف الكلمات الشائعة التي لا تحمل معنى دلاليًا كبيرًا.
- **الاشتقاق أو التصريف (Stemming / Lemmatization)**.**استخراج السمات**: تحويل النص إلى تمثيل رقمي يمكن للنموذج التعامل معه.
- **تطبيق نموذج التعلّم الآلي أو العميق**.

8.5 تقنيات و نماذج معالجة اللغات الطبيعية

تعتمد معالجة اللغات الطبيعية على عدة تقنيات، منها:

- نماذج إحصائية وتقليدية مثل Bag of Words وTF-IDF
- خوارزميات التعلّم الآلي
- الشبكات العصبية العميقة
- نماذج المحوّلات (Transformers) مثل BERT وGPT

و من افضلها ال BERT و الذي قمنا بالفعل باستخدامه في مشروعنا.

9.5 تمثيل الكلمات Embedding

يُقصد بالـ **Embedding** تحويل الكلمات أو النصوص إلى تمثيل رقمي (متجهات عددية) يحافظ على المعنى الدلالي والعلاقات بين الكلمات. يتيح هذا الأسلوب للنماذج الحاسوبية فهم التشابه بين الكلمات وسياق استخدامها، حيث تكون الكلمات المتشابهة معنويًا قريبة من بعضها في الفضاء العددي. ويُعد الـ Embedding أساسًا مهمًا في معظم تطبيقات معالجة اللغات الطبيعية الحديثة.

10.5 تمثيل النص باستخدام TF-IDF

تُعد **(TF-IDF (Term Frequency – Inverse Document Frequency** إحدى الطرق التقليدية في تمثيل النصوص، حيث تعتمد على حساب أهمية الكلمة داخل مستند مقارنةً ببقية المستندات. تزداد قيمة الكلمة كلما تكررت في مستند معين وقلت في باقي المستندات. تُستخدم TF-IDF على نطاق واسع في تصنيف النصوص واسترجاع المعلومات، إلا أنها لا تأخذ السياق أو المعنى الدلالي للكلمات بعين الاعتبار.

11.5 نموذج BERT

يُقصد بالـ **Embedding** تحويل الكلمات أو النصوص إلى تمثيل رقمي (متجهات عددية) يحافظ على المعنى الدلالي والعلاقات بين الكلمات. يتيح هذا الأسلوب للنماذج الحاسوبية فهم التشابه بين الكلمات وسياق استخدامها، حيث تكون الكلمات المتشابهة معنويًا قريبة من بعضها في الفضاء العددي. ويُعد الـ Embedding أساسًا مهمًا في معظم تطبيقات معالجة اللغات الطبيعية الحديث

6- منهجية العمل

1. البحث الدلالي Symantec Search

1.1.6 مقدمة

يهدف هذا الجزء إلى بناء **نظام بحث دلالي (Semantic Search)** على بيانات منتجات الأزياء من أمازون، بحيث يمكن للمستخدم البحث باستخدام وصف نصي (مثل: "black leather jacket for men")، والنظام يعيد المنتجات الأكثر تطابقاً **دلالياً**، وليس فقط بالكلمات المفتاحية التقليدية.

2.1.6 جمع البيانات

تم استخدام بيانات **Amazon Fashion Dataset** من amazon نفسها والتي تحتوي على:

- بيانات المنتجات: title, brand, description, feature, details, fit, imageURL.
- تقييمات العملاء: reviewText, summary, overall.
- بعض الحقول الأخرى مثل also_buy, also_view لم تُستخدم لأنها لا تضيف قيمة كبيرة للنظام الدلالي.

3.1.6 مراحل المعالجة المسبقة Pre-processing

شملت الخطوات التالية:

- تنظيف النصوص من HTML وURLs والعلامات غير الضرورية.
- توحيد الحروف إلى صغيرة (lowercase) وحذف الفراغات الزائدة.
- دمج الحقول المختلفة لكل منتج في مستند واحد يمثل المنتج بشكل كامل، بحيث يحتوي على: Title | Brand | Description | Features | Details | Fit | Review Summary | Review Text.
- التعامل مع القيم العددية الناقصة (مثل تقييم المنتج) باستبدالها بمتوسط التقييم.

الغرض من هذه المرحلة هو تجهيز النصوص **لتحويلها إلى تمثيلات عددية (embeddings)** لاحقاً.

4.1.6 تحويل النصوص الى تمثيلات عددية **Embedding**

تم استخدام نموذج حديث **SentenceTransformer**، بالتحديد النموذج:
BAAI/bge-large-en-v1.5
والذي يعتمد على **تقنية Transformer الحديثة** لتحويل النصوص إلى متجهات عددية عالية الأبعاد تعكس المعنى الدلالي للنصوص.

- تم تقسيم البيانات إلى **Chunks** صغيرة لتسريع العملية وتجنب نفاد الذاكرة.
- لكل جزء (Chunk) تم إنشاء **تمثيلات عددية مدمجة** وتم حفظها على القرص، حتى يمكن استكمال المعالجة إذا توقف النظام.
- تم تطبيع التمثيلات (Normalize) لضمان أن كل متجه له طول واحد، لتسهيل حساب التشابه بين المنتجات.

5.1.6 تطبيق البحث الدلالي

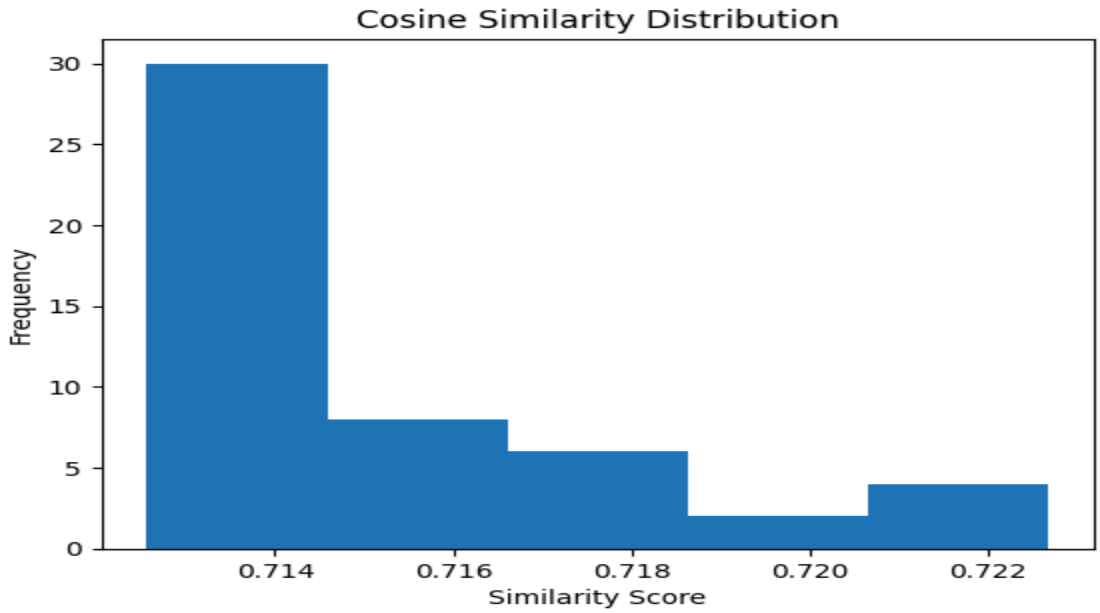
عند استلام استعلام المستخدم، يتم تحويله إلى **تمثيل عددي** باستخدام نفس النموذج.
ثم يتم حساب **Cosine Similarity** بين تمثيل الاستعلام وتمثيلات كل المنتجات.

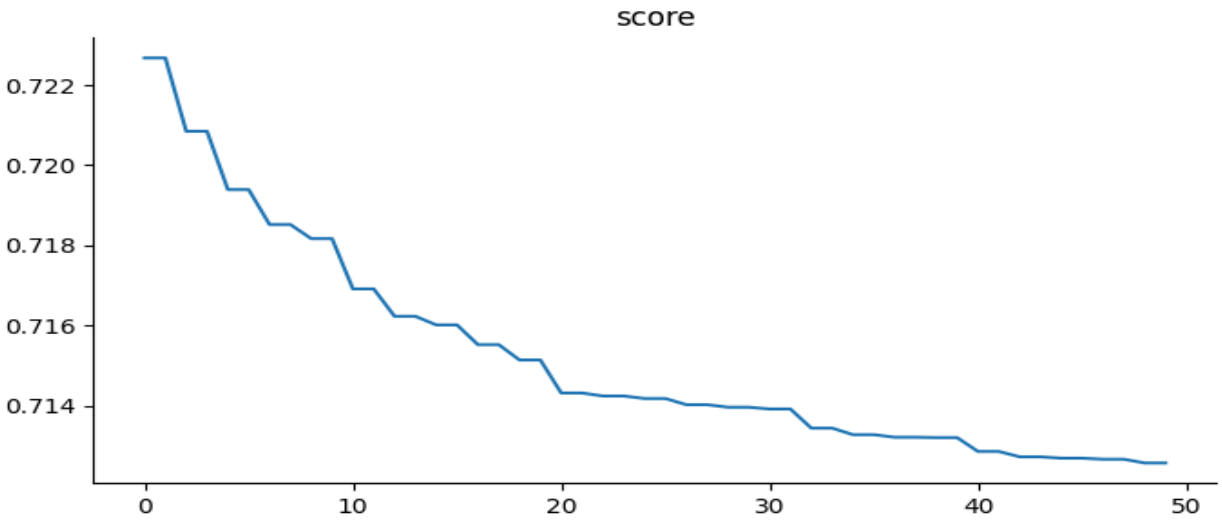
- أعلى القيم تمثل المنتجات الأكثر تطابقاً دلالياً مع الاستعلام.
- النتائج تُعرض مع **الاسم، العلامة التجارية، ودرجة التشابه**.

6.1.6 النتائج

تم استخدام **توزيع درجات التشابه (Cosine Similarity)** لرؤية مدى توافق النتائج:

- المحور X يمثل **درجة التشابه** بين الاستعلام والمنتجات.
- المحور Y يمثل **عدد المنتجات** التي حصلت على كل درجة تشابه.
- معظم النتائج تقع في نطاق عالي (0.7<)، مما يدل على جودة النظام في تحديد المنتجات الأقرب دلالياً





هذه كانت احد التجارب للحصول على 50 نتيجة بحث عن "black leather jacket for men"

2. نظام التوصية الهجين

1.2.6 مقدمة

يهدف هذا الجزء إلى بناء **نظام توصية هجين** يجمع بين أسلوبين رئيسيين: **التوصية المبنية على المحتوى (Content-Based)** و **التصفية التعاونية (Collaborative)**

Filtering)، بحيث يمكن للمستخدم الحصول على منتجات تتوافق مع تفضيلاته الشخصية وتجارب مستخدمين مشابهين.

النظام يحسب لكل منتج **درجة توافق مركبة (hybrid_score)**، والتي تعكس احتمالية إعجاب المستخدم بالمنتج، ويتم ترتيب النتائج بناءً على هذه الدرجة.

2.2.6 جمع البيانات

تم استخدام نفس بيانات **Amazon Fashion Dataset** التي تحتوي على:

- **بيانات المنتج:** title, brand, description.
- **تقييمات العملاء:** reviewerID, asin, overall.

3.2.6 مراحل المعالجة المسبقة **Pre-processing**

شملت الخطوات التالية:

- تنظيف النصوص من HTML وURLs والعلامات غير الضرورية.
- توحيد الحروف إلى صغيرة (lowercase) وحذف الفراغات الزائدة.
- دمج الحقول النصية لكل منتج مع إضافة **أوزان مختلفة** لكل حقل في نموذج المحتوى:
  - العنوان (Title) وزن 3
  - العلامة التجارية (Brand) وزن 2
  - الوصف (Description) وزن 1
- التعامل مع القيم العددية الناقصة في تقييمات العملاء (overall) باستبدالها بمتوسط التقييم.

الهدف من هذه المرحلة هو تجهيز البيانات **لبناء نموذج Content-Based وتحليل Collaborative Filtering** لاحقًا.

4.2.6 الوصية المبنية على المحتوى **Content based**

يعتمد هذا الجزء على **تشابه المنتجات من حيث المحتوى النصي**:

- يتم تحويل النصوص لكل منتج إلى **تمثيلات عددية** باستخدام أسلوب TF-IDF.

- لكل مستخدم يتم بناء **ملف شخصي (User Profile)** يدمج المنتجات التي قيمها سابقًا، مع مراعاة الأوزان المختلفة للحقول النصية.
- يتم حساب **Cosine Similarity** بين ملف المستخدم والمنتجات الأخرى لتحديد المنتجات الأكثر توافقًا.

5.2.6 التصفية الدلالية Collaborative Filtering

يعتمد هذا الجزء على **تجارب المستخدمين الآخرين**:

- يتم استخدام تقييمات المستخدمين (overall) لبناء نموذج يتنبأ بتقييم المستخدم لكل منتج لم يقيمه بعد.
- الهدف هو استغلال **الأنماط المشتركة بين المستخدمين** لتقديم توصيات دقيقة.

6.2.6 النظام الهجين Hybrid recommendation system

- يتم دمج نتائج التوصية المبنية على المحتوى ونتائج التصفية التعاونية للحصول على **hybrid_score** لكل منتج:

$$CB \cdot (\alpha - 1) + CF \cdot \alpha = hybrid\_score$$

- α هو وزن التصفية التعاونية (على سبيل المثال 0.7)، وCB تمثل درجة التشابه المبنية على المحتوى، مطبقة على نفس نطاق التقييمات (1-5).
- يتم ترتيب المنتجات وفقًا لأعلى hybrid_score لعرض النتائج الأكثر احتمالاً لإرضاء المستخدم.

7.2.6 النتائج

تم الحصول على قائمة منتجات مرتبة لكل مستخدم وفق hybrid_score.

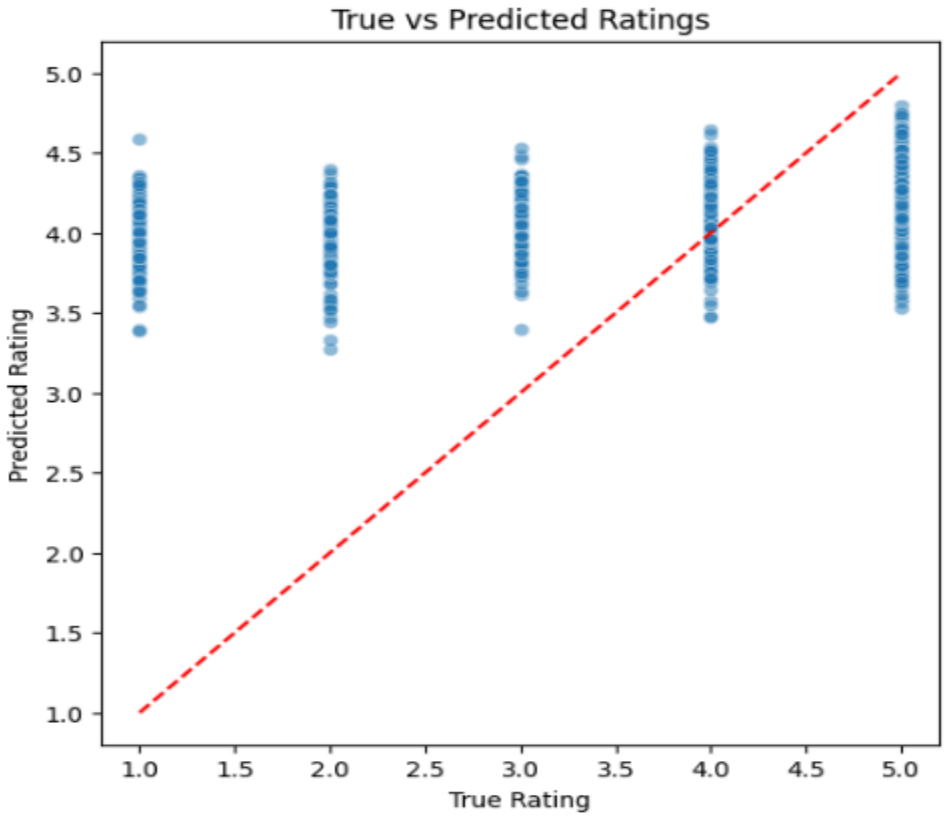
نطاق hybrid_score بين 1 و5، حيث القيم الأعلى تعني توافقًا أكبر مع تفضيلات المستخدم.

يمكن استخدام مخططات بيانية لاحقًا لعرض:

- توزيع hybrid_score لكل المستخدمين والمنتجات.
- مقارنة درجات CF وCB مع hybrid_score لتوضيح تأثير الدمج.

بالنسبة لل **Content - based** فقد حصلنا على نتائج مرضية بعد عدة تجارب حتى قمنا باضافة الاوزان لكل من [title, description, brand], حيث كانت نسبة التشابه باستخدام consine similarity تصل بين [0.32, 0.51] و هي نتائج ضعيفة, لكن بعد اضافة الاوزان توصلنا لنتائج اكثر ارضاءً [0.83, 0.923]

بالنسبة لل **Collaborative Filtering** فقد حصلنا على نتيجة جيدة عموماً حيث كانت نسبة الخطأ وفق معيار **RMSE** هو **1.09** و ذلك باستخدام **SVD**



اما النتائج النهائية بعد دمج كلاهما في ال **Hybrid recommendation system**
حصلنا على نتائج مرضية [3.641, 3.385] و ذلك لكل 10 نتائج

References لمراجع

[1] Strub, Florian, Romaric Gaudel, and Jérémie Mary.
"Hybrid Recommender System Based on Autoencoders."
In Proceedings of the Workshop on Deep Learning for Recommender Systems, pp. 11–16. ACM, 2016.

[2] Chen, Chao, Min Zhang, Yiqun Liu, Shaoping Ma, and Kuan Zhang.
"Variational Bandwidth Auto-Encoder for Hybrid Recommender Systems."
IEEE Transactions on Knowledge and Data Engineering 34, no. 11 (2022): 5405–5418.

[3] Alhijawi, Bushra, and Emad Abu-Shanab.
"Hybrid Recommendation by Incorporating the Sentiment of Product Reviews."
Information Sciences 608 (2022): 35–52.

[4] He, Xiangnan, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua.
"Neural Collaborative Filtering."
In Proceedings of the 26th International World Wide Web Conference (WWW), pp. 173–182. 2017
.
[5] Zhou, Kun, Yonghua Zhu, Yao Yu, Jialie Shen, and Jingbo Zhu.
"Web-Scale Semantic Product Search with Large Language Models."
In Proceedings of the European Conference on Information Retrieval (ECIR), pp. 75–90. Springer, 2023.

[6] Mao, Yuning, Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen.
"Lexically-Accelerated Dense Retrieval."
arXiv preprint arXiv:2307.16779 (2023).

[7] Yang, Yi, Yizhong Wang, Kai Zhang, and Zhiyuan Liu.
"Semantic Retrieval at Walmart."
arXiv preprint arXiv:2412.04637 (2024)
.
[8] Radford, Alec, Jong Wook Kim, Chris Hallacy, et al.
"Learning Transferable Visual Models From Natural Language Supervision."
In Proceedings of the International Conference on Machine Learning (ICML), pp. 8748–8763. 2021.

[9] Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes.
"Multimodal Semantic Retrieval for Product Search."
arXiv preprint arXiv:2501.07365 (2025).

[10] McAuley, Julian, and Jure Leskovec.
"From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise through Online Reviews."
In Proceedings of the 22nd International World Wide Web Conference (WWW), pp. 897–908. 2013.

[11] Mitchell, Tom M., *Machine Learning*, McGraw-Hill, 1997.

[12] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[13] Jurafsky, Daniel, and James H. Martin, *Speech and Language Processing*, Pearson, 3rd Edition, 2023.

[14] Eisenstein, Jacob, *Introduction to Natural Language Processing*, MIT Press, 2019.

[15] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint arXiv:1301.3781, 2013.

[16] Pennington, Jeffrey, Richard Socher, and Christopher Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.

[17] Salton, Gerard, and Christopher Buckley, "Term-weighting Approaches in Automatic Text Retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[18] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[19] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL, pp. 4171–4186, 2019.

[20] Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al., "Attention Is All You Need," Proceedings of NeurIPS, pp. 5998–6008, 2017.