تمثيل استعلامات SQL باستخدام نموذج BERT

م.سوزان محمد التركماني د.مهند رجب

الملخص:

تحليل الاستعلامات المطبقة على قواعد البيانات يمثل نقطة الإنطلاق للعديد من المهام الرئيسية في إدارة قواعد البيانات، مثل تلخيص الأحمال، وضبط بالفهارس، ومهام إدارية أخرى. ونظراً لصعوبة إجراء تحليلات يدوية مع توسع أحجام قواعد البيانات ومارافقها من تعقيد في الاستعلامات وزيادة في حجم الأحمال المطبقة على القاعدة، كان من الضروري إيجاد تمثيلات رقمية فعالة للاستعلامات لتمثيل المعلومات بشكل دقيق يمكن استخدامها في خوارزميات التعلم الآلي، على سبيل المثال كإدخال لخوارزميات التعلم المعزز المصممة لضبط الفهارس.

تهدف هذه الدراسة إلى استغلال نموذج لغوي متقدم، وهو نموذجBERT ، لإنشاء تمثيلات للاستعلامات. يشمل البحث استخراج التضمينات باستخدام النموذج، ثم تقييم كفاءتها في مهمتين اكتشاف التشابه بين الاستعلامات والتجميع.

أظهرت النتائج في مهمة اكتشاف التشابه الجودة العالية لتضمينات النموذج، حيث تفوق على أساليب تشابه الاستعلامات المرجعية في جميع مجموعات بيانات الإختبار بالنسبة لمعامل (Silhouette)، ووصل لتحسين لا يقل عن 90% بالنسبة لهذا المقياس كما حقق انخفاضاً كبيراً في معامل التماسك (BetaCV) بنسبة لا تقل عن 82% مقارنةً بأساليب التشابه المرجعية.

تم أيضاً تطبيق اختبارات على مستوى التجميع Clustering، حيث تمت مقارنة نموذج BERTمع ثلاثة نماذج لغوية معروفة أخرى باستخدام ثلاث خوارزميات تجميع و أظهرت

النتائج تفوق نموذجBERT بشكل ملحوظ في تشكيل تجمعات متماسكة ومنفصلة عبر مختلف مجموعات البيانات، خاصةً في خوارزميات Kmeans وHAC مع أداء جيد في خوارزمية Optics.

Representing SQL Queries using the Bert Model

Abstract:

uery analysis applied to databases is a starting point for many critical tasks in database management, such as workload summarization, index tuning, and other administrative tasks. Given the difficulty of performing manual analyses with the expansion of database sizes, the complexity of queries, and the increase in workload, it was essential to find effective numerical representations of queries to accurately represent information for use in machine learning algorithms. For example, these representations can serve as input for reinforcement learning algorithms designed for index tuning.

This study aims to utilize an advanced language model, BERT, to create representations for queries. The research involves extracting embeddings using the model and then evaluating their effectiveness in two tasks: similarity detection between queries and clustering.

The results showed high-quality embeddings for the model in the similarity detection task. BERT outperformed reference query similarity methods in all test datasets in terms of the silhouette score, achieving an improvement of at least 90% for this metric,

and significantly reducing the BetaCV coherence measure by at least 82% compared to reference similarity methods.

Clustering-level tests were also conducted, where BERT was compared with three other well-known language models using three clustering algorithms. The results demonstrated BERT's notable superiority in forming cohesive and distinct clusters across various datasets, especially with K-means and HAC algorithms, while also performing well with the Optics algorithm.

Keywords: Query Embeddings, Workload Analysis, Natural Language Processing, BERT Model, Similarity Metrics, Clustering.

مقدمة:

تُعتبر قواعد البيانات من الركائز الأساسية في الأنظمة الحديثة، حيث تلعب دورًا حيويًا في تنظيم وإدارة البيانات بكفاءة. تمثل قواعد البيانات محوراً رئيسياً في اتخاذ القرارات للأعمال والمؤسسات، ولها تأثير كبير على دقة التحليلات وأداء النظام بشكل عام. مع زيادة النشاط الرقمي وتطور التكنولوجيا، شهدت قواعد البيانات نمواً كبيراً في حجم البيانات المخزنة، مما أدى إلى تعقيد تنفيذ الاستعلامات وتحديات إضافية في إدارة الأحمال المطبقة على القواعد.

تحليل الاستعلامات يُعد من المهام الأساسية في إدارة قواعد البيانات، إذ يمكن أن يُحسن أداء النظام بشكل كبير من خلال ضبط الفهارس، تلخيص الأحمال، وغيرها من المهام الإدارية. ومع التزايد في حجم وتعقيد الاستعلامات، أصبحت الأساليب اليدوية لتحليل هذه الاستعلامات غير كافية. لذا، كان من الضروري البحث عن تمثيلات رقمية فعالة للاستعلامات يمكن استخدامها في خوارزميات التعلم الآلي لتحسين الأداء.

في هذا السياق، يتفوق نموذج BERT (Bidirectional Encoder Representations from Transformers)، وهو نموذج لغوي متقدم يعتمد على معمارية المحولات، في معالجة النصوص وفهم السياق. نظراً لقدرته على التقاط التمثيلات

السياقية بدقة، فإن BERT يمثل أداة واعدة لتمثيل استعلامات SQL وتحليل الأحمال المطبقة على قواعد البيانات.

تسعى هذه الدراسة إلى استغلال نموذج BERT لإنشاء تمثيلات للاستعلامات SQL وتقييم كفاءتها في مهمتين رئيسيتين: اكتشاف التشابه بين الاستعلامات والتجميع. تتضمن الدراسة استخراج التضمينات باستخدام نموذجBERT ، ومن ثم تقييم جودة هذه التضمينات مقارنة بأساليب تشابه الاستعلامات التقليدية.

أظهرت نتائج الدراسة في مهمة اكتشاف التشابه أن تضمينات BERT تتفوق بشكل ملحوظ على الأساليب المرجعية في جميع مجموعات بيانات الاختبار، حيث حققت تحسيناً لا يقل عن 90% في معامل (Silhouette) وانخفاضاً كبيراً بنسبة لا تقل عن 82% في معامل التماسك .(BetaCV) كما تم تطبيق اختبارات على مستوى التجميع، حيث تمت مقارنة نموذج BERT مع ثلاثة نماذج لغوية أخرى باستخدام ثلاث خوارزميات تجميع. أظهرت النتائج أن BERT يتفوق بشكل كبير في تشكيل تجمعات متماسكة ومنفصلة، خاصةً في خوارزميات Kmeans و HAC، مع أداء جيد في خوارزميةOptics .

من خلال هذه الدراسة، نهدف إلى تحسين فهمنا لفعالية نموذج BERT في تحليل استعلامات SQL وتقديم رؤى قيمة لتحسين أداء قواعد البيانات

1. أهداف البحث:

الهدف الرئيسي لهذه الدراسة هو تقييم كفاءة و فعالية تضمينات BERT في تمثيل المعلومات الموجودة في استعلامات SQL حيث قمنا بهذا التقييم من خلال استخدام هذه التضمينات في مهمة التقاط التشابه الكامن بين الاستعلامات حيث أن الاستعلامات المتشابهة قد يتم تنفيذها لأداء واجبات متماثلة فعلى سبيل المثال:

1. select distinct course.course_id, course.title from course
2. select course_id, title from course

استعلامان متشابهان على الرغم من بعض الاختلاف في صياغتهما إلا أنهما يقومان باسترجاع نفس السجلات وبالتالي يؤديان إلى نفس النتائج.

يأتي اختيارنا لنموذج BERT استناداً إلى القدرة المميزة لهذا النموذج على احتساب التمثيلات السياقية للبيانات النصية، وللقيام بهذا التقييم نقوم بإجراء تجارب على عدة مجموعات بيانات تحوي استعلامات متعددة مطبقة على قواعد بيانات مختلفة، والتي تمثل أنواع متنوعة من الأحمال.

2. الدراسات السابقة:

تقدم الورقة البحثية [10] نموذجاً مبتكراً يعتمد على تقنية BERT لتجميع أكواد البرمجة بناءً على وظائفها. باستخدام CuBERT، وهو نموذج BERT مُدرب مسبقاً على أكواد البرمجة، يتفوق COCLUBERT بشكل كبير على النماذج التقليدية في دقة التجميع، حيث يحسن مؤشرات الأداء بشكل ملحوظ مثل مؤشر Dunn الذي زاد بمقدار 141 ضعفاً، ومؤشر Silhouette الذي تضاعف مرتين، ومؤشر Adjusted Rand الذي زاد بمقدار 11 ضعفاً. يبرز النموذج بفضل تقنيات تحسين متعددة مثل Triplet Loss و Unsupervised Clustering Loss و Deep Robust Clustering (DRC) Loss. ومع ذلك، يتطلب النموذج موارد حسابية كبيرة للتدريب والتنفيذ، مما يجعله معقداً ويحد من تطبيقه في بيئات أقل قوة. بالإضافة إلى ذلك، يعتمد على تحليل أسماء الدوال لتحديد الوظائف المشابهة، مما قد يكون محدوداً في بعض الحالات.

يستعرض الباحثون في [25] تأثير التعديلات على ميزات استعلامات SQL في تصنيفها كخبيثة أو غير خبيثة باستخدام نموذج Random Forest، من خلال تحليل تأثير التعديلات الفردية والمجمعة على الميزات. تركز الدراسة على مجموعة من الميزات الرئيسية منها : ميزة وجود أحرف خاصة: هذه الميزة تشير إلى وجود أحرف خاصة مثل '--' أو 'OR' في الاستعلامات، والتي غالباً ما تُستخدم في هجمات SQL injection. تفعيل هذه الميزة حول 87.62% من الاستعلامات غير الخبيثة إلى خبيثة، مما يبرز حساسيتها للعلامات التي قد تدل على هجوم.

ميزة وجود الكلمة المفتاحية ORDER: تتعلق بوجود كلمة ORDER التي تُستخدم لتنظيم البيانات. تفعيل هذه الميزة ساعد في تصنيف 11.25% من الاستعلامات الخبيثة كغير خبيثة، مما يشير إلى أن ORDER غالباً ما يرتبط بالاستعلامات المشروعة.

ميزة غياب الكلمة المفتاحية FROM: تتعلق بغياب كلمة FROM، وهي ضرورية لتحديد الجدول في استعلامات SQL. غيابها حول 7.88% من الاستعلامات غير الخبيثة إلى خبيثة، مما يشير إلى أهمية وجود هذه العناصر الأساسية.

الميزة غياب رمز =: تشير إلى غياب رمز = الذي يُستخدم للمقارنات في SQL. تفعيل هذه الميزة أثر على 9.54% من الاستعلامات الخبيثة بتحويلها إلى غير خبيثة، مما يبرز دور الرموز الأساسية في تحديد نوايا الاستعلامات.

توصلت الدراسة الى أن التعديلات البسيطة على ميزات محددة يمكن أن تؤدي إلى تغييرات كبيرة في تصنيف الاستعلامات، حيث تكشف الدراسة عن حساسية النموذج للتعديلات على الميزات. مما يسلط الضوء على أهمية بعض الميزات في تصنيف الاستعلامات بشكل صحيح. وعلى الرغم من أن الدراسة تناولت تأثيرات تعديل الميزات بشكل فردي، فإن التفاعل بين الميزات قد يكون معقداً وقد يتطلب تحليلاً أعمق.

يقترح الباحثون في [15] مقياساً زوجياً للتشابه بين استعلامي SQL وعلى الرغم من أنه ليس الهدف الأساسي لبحثهم حيث أنهم يهدفون إلى تحسين اختيار المناظير في مستودعات البيانات اعتماداً على الاستعلامات الموجهة إلى النظام. قاموا بداية باستخراج السمات الممثلة لكل استعلام ضمن الحمل، حيث أن السمات التمثيلية هي تلك السمات التي تظهر في شرط Where وفي عبارات Group by وقاموا ببناء مصفوفة كل سطر فيها يمثل استعلام qi و العمود يعبر عن سمة تمثيلية aj حيث يتم اضافة أعمدة لجميع السمات التمثيلية لكافة الاستعلامات الموجودة ضمن الحمل قيمة كل خانة cij ضمن هذه المصفوفة يمكن أن تكون 1 أو 0 وتعبر عن اذا ما كانت الصفة aj موجودة أم غير

موجود في الاستعلام qi على الترتيب، أثناء إنشاء المصفوفة، لا يهم إذا كان العنصر يظهر أكثر من مرة أو مكان العنصر. في الخطوة التالية يتم تطبيق مقياس التشابه المقترح حيث:

من أجل إيجاد التشابه بين الاستعلامين qk، ql يتم مقارنة السطرين المعبرين عنهما ضمن المصفوفة السابقة ويتم استخدام مسافة هامينغ (Hamming distance) لإيجاد عدد السمات المشتركة بين الاستعلامين وكون هدف البحث الأساسي ليس إيجاد التشابه بين الاستعلام وإنما كما ذكرنا سابقا هو تحسين اختيار المناظير في مستودعات البيانات فلم يتم ذكر نتائج مخصصة لفعالية اسلوب التشابه المقترح أو لمقارنته مع اساليب أخرى أو حتى نسبة الاستعلامات المتشابهة التي تم الحصول عليها باستخدام المقياس.

تتبنى الدراسة [16] تحليل سجل الاستعلام بغاية تحليل الأحمال المطبقة، وتقدم مجموعة من التجارب على مجموعة بيانات Sloan Digital Sky Survey (SDSS) . تنقيب النصوص، أو اكتشاف المعرفة في النصوص Knowledge Discovery in (KDT) Text، هو تمديد لعملية اكتشاف المعرفة التقليدية في قواعد البيانات Knowledge Discovery in Databases (KDD) ، وهي عملية تهتم بتحديد الأنماط الصحيحة والتي من المرجح أن تكون مفيدة في البيانات، ولكنه يستهدف البيانات غير المهيكلة أو شبه المهيكلة بدلاً من قواعد البيانات العادية، مثل رسائل البريد الإلكتروني والمستندات النصية الكاملة وبعض أنواع الملفات(على سبيل المثال، HTML و XML). إنها مجال متعدد التخصصات يمكن أن يشمل أمور أخرى كاسترجاع المعلومات، وتعلم الآلة، ومعالجة اللغة الطبيعية. تم التعامل مع استعلامات SQL في هذا السياق على أنها وثائق صغيرة حيث يمكن الاستفادة من الهيكل الذي توفره اللغة لتحسين خطوة معالجة الاستعلامات لتناسب الحالات الخاصة الموجودة. على سبيل المثال، لا حاجة لإزالة كلمات التوقف. يتم استخراج المصطلحات في عمليات from، join، select، group-by، order-by على حدة ويتم تسجيل تكرار ظهورها في متجه ميزة ويستخدمونه لحساب التشابه التدريجي بين الاستعلامات باستخدام تشابه جيب التمام Cosine Similarity.

سلطت الدراسة [2] الضوء على أهمية أساليب التشابه وفعالية تقنيات هندسة الميزات في تحسين دقة تجميع الاستعلامات حيث تم التركيز على إيجاد التشابه بين الإستعلامات كون الإستعلامات التي لها بنية متشابهة تشير إلى أنه قد يتم تنفيذها لأداء واجبات مماثلة.

تم تقييم ثلاثة أساليب استدلالية لتشابه الاستعلامات المقترحة في الأدبيات المطبقة على تجميع الاستعلامات (Aligon، Aouiche، Makiyama) المذكورة في [13]، [15] و[16] كون هذه الأساليب تركز على البنية النحوية للاستعلام ولا تتطلب الوصول إلى البيانات في قاعدة البيانات حيث أن متطلبات الوصول إلى البيانات تتسبب في انتهاك للخصوصية، لذلك فقد تم التركيز على النهج القائم على البنية النحوية.

بعد الانتهاء من التقييم تم تطبيق تقنية هندسة الميزات المسماة التنظيم (التي تم اقتراحها كخطوة إضافية من قبل الباحثين) لتحسين دقة تجميع الاستعلامات. تستخدم الدراسة ثلاث مجموعات بيانات – استعلامات مؤلفة من أجوبة الطلاب من جامعةIIT Bombay ، واستعلامات تشكل اجابات الطلاب من جامعة بوفالو، وسجلات SQL من تطبيق Google+. ويظهر التقييم للأساليب الثلاثة أن Aligon يقوم بأداء أفضل عبر المجموعات الثلاث بناءً على مجموعة من مقاييس الأداء الخاصة بعملية التجميع مثل Silhouette Coefficient، Betacv وDunn Index . ومع ذلك، لا تؤدي أي من الأساليب بشكل جيد كما هو مرغوب، فقام الباحثون باقتراح وتقييم خطوة معالجة مسبقة تتضمن إنشاء تمثيل أكثر فعالية للإستعلامات بالإستفادة من قواعد مكافئة الإستعلام وعمليات تجزئة البيانات. حسنت خطوة المعالجة المقترحة أداء الأساليب بشكل واضح، خاصة من حيث Silhouette Coefficient حيث تراوحت نسبة التحسين بين 18% في بعض مجموعات الاختبار ووصلت حتى 200% في مجموعات أخرى كما حسنت الأداء على مستوى مقياس Betacv بشكل واضح في بعض مجموعات البيانات حيث وصلت نسبة التحسين إلى 33% فيما لم تؤثر بشكل كبير على الأداء في مجموعات أخرى كمجموعة بيانات PocketData–Google.

أما في [14] فقد كان تركيز الباحثين على تصنيف لغات التخصص بالمجال (DSL)، وبشكل خاص استعلامات SQL. بهذا الصدد، تم تقديم TRANSQL واستخدام نموذجي لغة، BERT و GPT-3، لايجاد تمثيل لاستعلامات SQL وتصنيفها. أكدت النتائج أن TRANSQL القائم على BERT يحتاج إلى كميات كبيرة من البيانات في خطوة ضبط النموذج، وكون مجموعة البيانات التي تم العمل عليها كانت تحوي عدد قليل من الاستعلامات لذلك لم يتمكن النموذج من أن يؤدي بشكل جيد في سياق التعلم القليل (few_shot). و نظراً لأن GPT-3 هو نموذج فعال في سياق التعلم القليل، يتفوق TRANSQL القائم على GPT على TRANSQL القائم على BERT. علاوة على ذلك، كشفت التجارب أن استراتيجية اختيار الأمثلة في السياق في TRANSQL القائم على GPT تلعب دوراً حاسماً في أداء مهمة التصنيف، حيث أدت إلى تحسين يصل إلى 33% في أداء المهمة.

3. تضمين الكلمات Word Embedding :

تضمين الكلمات في معالجة اللغة الطبيعية هو مصطلح مهم يُعبر عن تمثيل الكلمات لتحليل النصوص في شكل متجهات ذات قيم حقيقية.

بعض نماذج تضمين الكلمات:

:Word2Vec

يمثل Word2Vec [22] إحدى أولى النماذج التي قامت بتقديم فكرة تضمين الكلمات. تستخدم Word2Vec نموذجاً سهل الاستخدام يعتمد على الشبكات العصبونيّة.

يعمل النموذج على تمثيل الكلمة اعتماداً على سياقها في متّجه أبعاده N، تمّ تطوير هذا الأسلوب من قبل توماس ميكولوف Tomas Mikolov في العام 2013 لجعل التّدريب في الشبكات العصبونيّة مبنيّ على التّضمين، والذي يحقق أكثر كفاءة ومنذ ذلك الحين أصبح المعيار الواقعيّ لتطوير تضمين الكلمة.

يتم تدريب النموذج باستخدام واحدة من طريقتين أساسيتين:

1. Continuous Bag of Words (CBOW): يحاول هذا النموذج التنبؤ بالكلمة المستهدفة بناءً على الكلمات السياقية المحيطة بها.
2. Skip-gram: في هذا النموذج، يتم تدريب النموذج على التنبؤ بالكلمات السياقية بناءً على الكلمة المستهدفة.

GloVe (Global Vectors for Word Representation)

GloVe [23]هو نموذج تضمين الكلمات طوره باحثون في مختبر الذكاء الصناعي بجامعة ستانفورد في عام 2014. يعتمد GloVe على تحليل المصفوفات المشتركة للكلمات في نصوص ضخمة. حيث يتم بناء مصفوفة كبيرة تحتوي على تكرارات مشتركة للكلمات، حيث تمثل كل خلية فيها عدد مرات ظهور كلمتين معاً في سياقات معينة. وتم استخراج التضمينات عبر تحليل المصفوفة باستخدام تقنيات مثل تحليل القيمة المفردة Singular Value Decomposition (SVD) وتقنيات التحسين العشوائي.

Word2Vec و GloVe هما أمثلة على نماذج التضمين التي يتم تدريبها على مجموعات نصية كبيرة بناءً على إحصائيات التواجد المشترك لتوليد تمثيلات دلالية للكلمات. ومع ذلك، فإن هذه النماذج تولد تضمينات كلمات غير مرتبطة بالسياق، مما يعني أن التمثيل الذي تُنشؤه لكلمة ما لا يرتبط بمعنى الكلمة في تلك الجملة.

:XLNet

XLNet [24] هو نموذج حديث لمعالجة اللغة الطبيعية (NLP) تم تطويره من قبل باحثين في شركة Google AI و Carnegie Mellon University.

يعتمد على تقنية المحولات (Transformers) ويستخدم آلية مبتكرة تسمى نمذجة اللغة بالتباديل Permutation Language Modeling، حيث يقوم بتوليد كافة التباديل الممكنة لتسلسل الكلمات في الجملة، مما يسمح له بالنظر إلى كل كلمة في سياقات مختلفة ومتنوعة. هذا النهج يمكن النموذج من تجنب التحيزات الناتجة عن الترتيب الثابت للكلمات ويمنحه فهماً أعمق وأكثر شمولاً للنص. يتميز XLNet أيضاً بدمجه بين مزايا النماذج

التوليدية والتشفيرية الذاتية، مما يجعله فعالاً في كل من فهم وتوليد النصوص. كما يستخدم XLNet آلية ذاكرة لتتبع المعلومات الطويلة المدى في النصوص، مما يعزز قدرته على التعامل مع النصوص الأطول بشكل فعال. الأداء المتفوق لـ XLNet في مجموعة متنوعة من مهام معالجة اللغة الطبيعية، مثل التصنيف النصي، واستخراج الكيانات، وفهم القراءة، والترجمة الآلية، يجعله أداة قوية ومتعددة الاستخدامات.

نموذج Bert:

هو نموذج معالجة لغوية (NLP) مفتوح المصدر، تم تطويره من قبل باحثين في شركة جوجل في العام 2018 وتم تقديمه ضمن الورقة البحثية [7] وقد أحدثت هذه الورقة ضجة في مجتمع التعلم الآلي من خلال تحقيق نتائج متقدمة في مجموعة واسعة من مهام معالجة اللغة الطبيعية، بما في ذلك الإجابة على الأسئلة واستنتاج اللغة الطبيعية وغيرها حيث يعتبر BERT قوياً في فهم السياق اللغوي وتمثيل المعاني بشكل فعال.

في السابق، كان يمكن لنماذج اللغة الطبيعية قراءة النصوص بتسلسل – إما من اليسار إلى اليمين أو من اليمين إلى اليسار – ولكنها لم تكن قادرة على القيام بكليهما في نفس الوقت. يتميز BERT بقدرته على قراءة النص في الاتجاهين في نفس الوقت، وهو ما يعرف بالثنائية الاتجاهية حيث يمكن للنموذج فهم السياق من اليسار إلى اليمين ومن اليمين إلى اليسار.

تم تدريب BERT مسبقاً على بيانات ضخمة حيث استخدم الباحثون مجموعة بيانات (BooksCorpus) التي تحتوي على 800 مليون كلمة [8] و ويكيبيديا باللغة الإنجليزية التي تحتوي على 2500 مليون كلمة، كما تمت عملية التدريب بالاعتماد على مهمتين من مهام معالجة اللغة الطبيعية: نمذجة اللغة بالكلمات المخفية (MLM) و التنبؤ بالجملة التالية (NSP).

هدف تدريب نمذجة اللغة بالكلمات المخفية (MLM) هو إخفاء كلمة في جملة ومن ثم توجيه النموذج للتنبؤ بالكلمة التي تم إخفاؤها (تمييزها) استناداً إلى السياق المحيط بها.

وهدف تدريب التنبؤ بالجملة التالية (NSP) هو توجيه النموذج للتنبؤ بما إذا كان لدى جملتين ارتباط منطقي وتسلسلي أو إذا كانت علاقتهما عشوائية فقط.

يعتمد BERT على المحولات (Transformers)، وهي نموذج تعلم عميق حيث يكون كل عنصر من الإخراج متصلاً بكل عنصر من الإدخال، ويتم حساب الأوزان بينهم بشكل ديناميكي استناداً إلى اتصالاتهم. (في مجال معالجة اللغة الطبيعية، يطلق على هذه العملية اسم "attention").

نمذجة اللغة بالكلمات المخفية (MLM) :

هي تقنية تعلم عميق شهيرة تستخدم في مهام معالجة اللغة الطبيعية (NLP)، خاصة في تدريب نماذج النقل مثل BERT، GPT-2، و RoBERTa.

في MLM، يتم استبدال 15% من الكلمات في كل تسلسل برمز [MASK] قبل إدخال تسلسلات الكلمات إلى BERT ويتم تدريب النموذج على التنبؤ بالقيمة الأصلية للكلمات التي تم إخفاؤها استناداً إلى السياق المحيط به. الفكرة وراء ذلك هي تدريب النموذج على فهم سياق الكلمات وعلاقاتها مع كلمات أخرى في الجملة.

MLM هي تقنية تعلم ذاتي التوجيه، مما يعني أن النموذج يتعلم إنتاج النص بدون الحاجة إلى توجيهات صريحة أو تسميات، بل باستخدام النص الداخلي نفسه كإشراف. وهذا يجعلها أداة متعددة الاستخدام وفعالة لمجموعة واسعة من مهام NLP، بما في ذلك تصنيف النص، الإجابة على الأسئلة، وغيرها.

أثناء عملية التدريب، يتم تحديث النموذج استناداً إلى الفرق بين تنبؤاته والكلمات الفعلية في الجملة. تساعد هذه المرحلة النموذج على تعلم تمثيلات سياقية مفيدة للكلمات، يمكن بعد ذلك تعديلها بشكل دقيق لمهام معالجة اللغة الطبيعية المحددة. الفكرة وراء MLM هي الاستفادة من كميات كبيرة من بيانات النص المتاحة لتعلم نموذج لغوي متعدد الاستخدام يمكن تطبيقه على مشاكل مختلفة في مجال NLP.

تابع الخسارة المقدم في BERT يأخذ في عين الاعتبار فقط التنبؤ بقيم الرموز المخفية ويتجاهل التنبؤ ببقية الكلمات (غير المخفية). ونتيجة لذلك، يتقدم النموذج ببطء أكبر من النماذج الاتجاهية، ولكن في المقابل يكون لديه وعي أكبر بالسياق.

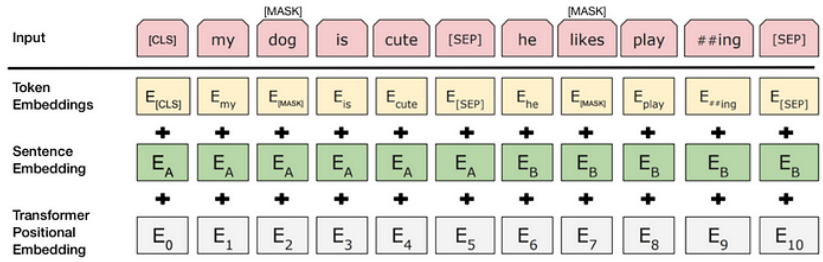
التنبؤ بالجملة التالية (NSP):

هي مهمة مسبقة التدريب تستخدم في تدريب نماذج BERT تم تصميم مهمة NSP لتمكين BERT من تعلم علاقات لا تقتصر فقط على الكلمات داخل الجملة، بل تشمل أيضًا العلاقات بين الجمل في وثيقة أو مجموعة نصوص.

في هذه المهمة، يأخذ BERT زوجاً من الجمل كإدخال أثناء مرحلة ما قبل التدريب. تستخدم هذه الجمل عادة من نصوص طويلة. يتواجد رمز فاصل خاص (مثل [SEP]) بين الجملتين في الإدخال. ويطلب منه تحديد ما إذا كانت الجملة الثانية تأتي بعد الجملة الأولى بشكل مباشر أم لا، أي أن مهمة NSP هي مهمة تصنيف ثنائي حيث يتم تدريب النموذج على تصنيفين ممكنين:

1. تصنيف (IsNext): إذا كانت الجملة الثانية تتبع بشكل منطقي الجملة الأولى، يتم تعيين التصنيف إلى "IsNext".
2. تصنيف (NotNext): إذا لم تكن الجملة الثانية هي الجملة التالية، يتم تعيين التصنيف إلى "NotNext".

تساعد هذه المهمة النموذج في التقاط العلاقات والتماسك بين الجمل في وثيقة. خلال عملية التدريب، 50% من المدخلات تكون عبارة عن أزواج ايجابية من الجمل أي أن الجملة الثانية هي الجملة التالية للجملة الأولى، و 50% يتم اختيار أزواج من جمل عشوائية من المجموعة النصية، يفترض أن تكون الجملة الثانية منفصلة عن الجملة الأولى. هذا يساعد النموذج على فهم العلاقات السياقية بين الجمل وتحسين قدرته على فهم السياق اللغوي والتنبؤ بالجمل التالية في النصوص الطويلة.



الشكل 1: تمثيل دخل Bert

يوضح الشكل 1 مكونات التضمينات لـ BERT وهي عبارة عن ثلاثة أجزاء:

- الرمز (Token): الرموز هي أساساً الكلمات. (يستخدم BERT مجموعة مفردات مكونة من حوالي 30 ألف رمز أما بالنسبة للكلمات لتي لا تظهر في مجموعة المفردات، يتم تجزئت هذه الكلمات إلى عدة رموز). الرمز الأول في التسلسل هو [CLS] الذي يكون مفيداً لمهام التصنيف. خلال التدريب السابق لـ MLM، يتم إخفاء بعض الرموز.
- الجملة (Sentence): تسلسل رموز الإدخال يمكن أن يكون قسماً واحداً أو قسمين. يعد القسم جزء متتالي من النص، وليس جملة لغوية فعلية. يتم إنهاء كل قسم برمز [SEP]، على سبيل المثال في الإجابة على الأسئلة، يكون السؤال هو القطاع الأول والإجابة هي القطاع الثاني.
- الموضع (Position): يمثل هذا المكون موضع الرمز ضمن التسلسل.

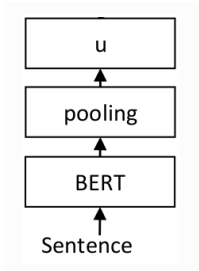
SBERT (Sentence-Bert):

عند تضمين الجمل أو النصوص، يكون الهدف إيجاد متجه كثيف ذو حجم ثابت انطاقاً من نص مدخل طوله متغير. صمم BERT بشكل أساسي لفهم المعاني على مستوى الكلمات word-level أو الرموز token-level داخل الجمل حيث يقدم النموذج تضمينات على مستوى الكلمة.

في الممارسة العملية، غالباً ما نحتاج إلى إنشاء تضمينات ليس لكلمات منفردة بل بدلاً من ذلك لجمل بأكملها، وبينما يُمكن ل BERT التقاط السياق بشكل جيد، إلا أنه لا ينمذج بشكل صريح العلاقات الدلالية بين الجمل بأكملها، لذلك ظهرت نسخة معدلة من نموذج BERT، مصممة خصيصاً لمهام التشابه الدلالي على مستوى الجملة وهو نموذج SBERT، يقوم SBERT بسد هذه الفجوة من خلال إنتاج تضمينات تمثل المعنى الدلالي العام للجمل، مما يتيح لها المقارنة بشكل أكثر دقة.

SBERT (Sentence-BERT) هو نموذج مصمم خصيصاً لإنتاج تضمينات معنوية للجمل. يقوم SBERT بتعديل هيكل BERT لإنشاء تضمينات على مستوى الجملة يمكن استخدامها لقياس التشابه الدلالي بين الجمل.

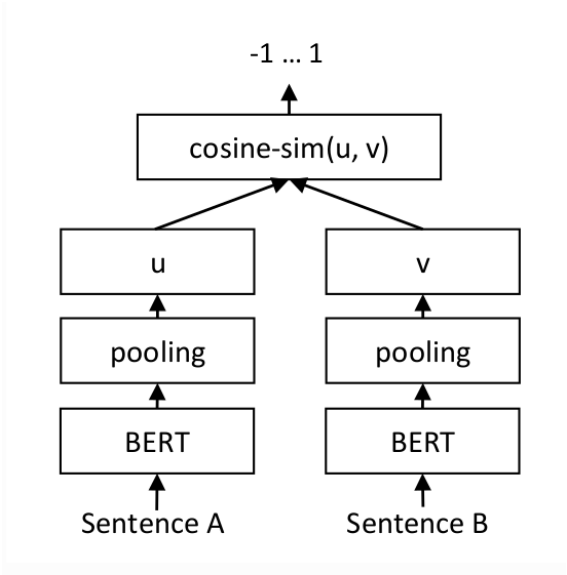
أبسط تصميم للشبكة العصبية الذي يمكن استخدامه هو التالي:



الشكل 2: بنية شبكة بسيطة لايجاد تضمينات للسلاسل النصية

يتم تغذية شبكة BERT بالجملة أو النص. تنتج شبكة BERT تضمينات متناسبة مع السياق لجميع الرموز المدخلة في النص. نظراً لأننا نريد تمثيلات ذات حجم ثابت (متجه u)، فإننا بحاجة إلى طبقة تجميع (Pooling). تتوفر خيارات تجميع مختلفة، وأبسطها هو التجميع الوسيط mean-pooling الذي يقوم بايجاد معدل جميع التضمينات (الكلمات) التي يقدمها BERT. هذا يعطي متجه في الخرج ذو أبعاد ثابتة بحجم 768 بحالة نموذج BERT-Base بغض النظر عن طول النص المدخل.

الشبكة السابقة، المكونة من طبقة BERT وطبقة التجميع، تشكل جزئية واحدة من نموذج SBERT. حيث تكون الشبكة كاملة بالشكل:



الشكل 3:بنية شبكة SBERT (Siamese–BERT)

هذا الشكل هو أحد النهج الشائعة والذي يعرف بالنموذج (Siamese-BERT) وهو تصميم يجمع بين نهج الشبكة المزدوجة (Siamese network) ونموذج BERT. يعتمد Siamese-BERT على استخدام شبكتين عصبيتين متطابقتين، كل منهما يأخذ إدخالاً مختلفاً (جمليتن) عند تمرير الجملتين عبر طبقات التجميع (يقترح مؤلفو SBERT اختيار طبقة متوسط التجميع mean-pooling كطبقة افتراضية) ، نحصل على متجهين بُعدهما 768 بحالة BERT-Base، يُرمزان بـ u و v. ويتم مقارنة تضميناتهما باستخدام مقياس مسافة مثل التشابه الجيبي. يكون هذا التصميم مفيداً بشكل خاص للمهام التي تتضمن مقارنات التشابه أو عدم التشابه بين المدخلات.

4. الضبط الدقيق لنموذج Bert:

تعني عملية "Fine-tuning" أو (الضبط الدقيق للنموذج) في سياق نماذج BERT ونماذج المحولات الأخرى المدربة مسبقاً استمرار تدريب النموذج على مجموعة بيانات

خاصة بمجال المهمة المراد استخدام النموذج لها بعد تدريبه المبدئي على مجموعة بيانات كبيرة وعامة. هذه العملية تجرى عادة لتكييف النموذج المدرب مسبقاً ليفهم بشكل أفضل الميزات الخاصة بمجموعة البيانات المرتبطة للمهمة المحددة.

يمكن استخدام BERT لمجموعة واسعة من المهام اللغوية، مع إضافة طبقة صغيرة إلى النموذج الأساسي:

- في مهام التصنيف مثلا يتم القيام بذلك بشكل مماثل لمهمة التنبؤ بالجملة التالية، عن طريق إضافة طبقة تصنيف.
- في التعرف على الكيانات المعترف بها (NER)، يتلقى البرنامج تسلسل نصي ويتعين عليه وضع علامات على مختلف أنواع الكيانات (شخص، منظمة، تاريخ، إلخ) التي تظهر في النص. باستخدام BERT، يمكن تدريب نموذج NER عن طريق تغذية الشعاع الناتج لكل رمز إلى طبقة تصنيف تتنبأ بتصنيف NER.

في تدريب الضبط الدقيق، تظل معظم المعلمات الفائقة كما هو الحال في تدريب BERT. قد قام فريق BERT باستخدام هذه التقنية لتحقيق نتائج متقدمة على مجموعة واسعة من المهام اللغة الطبيعية.

5. تقنيات التجميع clustering methods :

توجد العديد من تقنيات التجميع، ولكننا في هذا العمل نركز على التقنيات المستندة إلى المركز Centroid-based clustering ، الكثافة Density-based clustering، والتجميع الهرميHierarchical clustering .

التجميع المستند إلى المركز يقوم بتقسيم نقاط البيانات في مجموعة البيانات عن طريق إيجاد مراكز لعدد معين من المجموعات، ثم يتم تعيين كل نقطة بيانات لأقرب مركز مجموعة لها، بحيث يتم تقليل المسافات المربعة بين النقاط ومركز المجموعة إلى الحد الأدنى. تعد خوارزمية K-means [19] واحدة من أبرز الخوارزميات في هذا النوع.

التجميع المستند إلى الكثافة يعتمد على كثافة المناطق في البيانات لتشكيل المجموعات. الفكرة الرئيسية هي أن نقاط البيانات في المناطق عالية الكثافة تكون أكثر تشابهاً فيما بينها وتختلف عن تلك الموجودة في المناطق منخفضة الكثافة. من الأمثلة على هذه الطرق خوارزمية Ordering Points To Identify the Clustering Structure OPTICS [20]. وعلى عكس K-means، لا تتطلب هذه الخوارزميات تحديد عدد المجموعات مسبقاً.

التجميع الهرمي يشكل مجموعات باتباع بنية شجرية تستند إلى التسلسل الهرمي للبيانات، ويُنشئ مجموعات جديدة من المجموعات السابقة. التجميع الهرمي التراكمي (HAC) [21]، كأحد أمثلة هذا النوع من الخوارزميات، يبدأ بتعيين كل نقطة بيانات إلى مجموعة منفصلة، ثم يتم دمج المجموعات الأكثر تشابهاً تدريجياً لتشكيل مجموعات جديدة. بعد كل عملية دمج، يتم إعادة حساب مقاييس التشابه وتستمر عملية الدمج

6. الإطار العملي:

يتم في هذ القسم إجراء مقارنة في الأداء بين التمثيلات التي يتم الحصول عليها من نموذج Bert وأساليب التشابه السابقة المذكورة في الأدبيات إضافة الى معرفة كفاءة تضمينات Bert على مستوى التجميع clustering مقارنة بمجموعة من النماذج اللغوية الشهيرة.

مجموعات البيانات:

يتم تنفيذ مقارنة الأداء على عدة مجموعات مكونة من استعلامات:

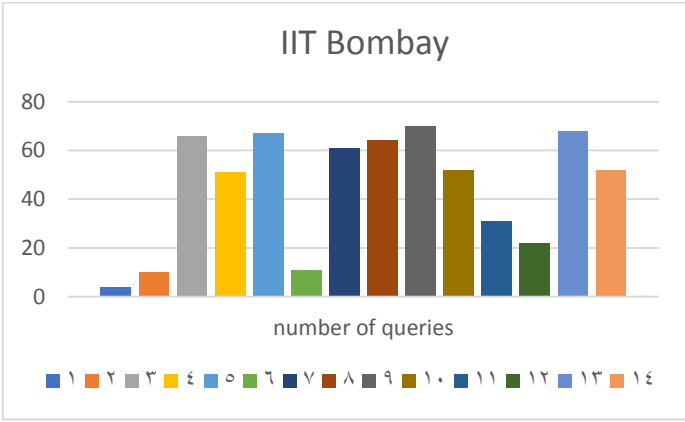
المجموعة الأولى IIT Bombay [17]، المستحصلة من جامعة IIT Bombay، تحتوي على إجابات الطلاب على امتحان في مادة قواعد البيانات على مدى سنتين. (14 سؤال).

المجموعتين الثانية UB Exam [2] والثالثة Goggle Plus [2] تم تقديمهما من خلال الورقة البحثية [2]، الأولى تتألف أيضا من إجابات الطلاب، وتتضمن اجابات لسؤالين تم

تنفيذهما كجزء من امتحانات منتصف الفصل لعامين متتاليين أما المجموعة الثانية، فهي تتضمن سجلات SQL تسجل جميع الأنشطة في قواعد البيانات على 11 هاتفاً ذكياً على مدى شهر واحد. يتم تصنيف هذه السجلات بواحدة من ثماني فئات مختلفة: الحساب، النشاط، التحليلات، جهات الاتصال، الأخبار، الصيانة، الوسائط، والصور.

المجموعة الأخيرة Automated-Grading-of-SQL-Statements [18] تم تقديمها من قبل الجامعة الوطنية الأسترالية تتضمن إجابات 393 طالبا على 15 تمرين.

مجموعة بيانات IIT Bombay :



الشكل 4: توزع الاستعلامات على الفئات ال14 ضمن مجموعة بيانات IIT Bombay

مثال للإستعلامات الموجودة في مجموعة البيانات:

label	query
1	select distinct course_id,title from course
1	select course_id,title from course
2	select course_id,title from course where dept_name='comp. sci.'
2	select distinct course.course_id,course.title from course where course.dept_name='comp. sci.'
2	select course.course_id,course.title from course where dept_name='comp. sci.'
2	select course id,title from course where course.dept name = 'comp. sci.'

الشكل 5: عينة من الاستعلامات في مجموعة بيانات IIT Bombay

مجموعة بيانات UB Exam:

السنة	السؤال
2014	How many distinct species of bird have ever been seen by the observer who saw the most birds on December 15، 2013?
2015	You are hired by a local birdwatching organization, who's database uses the Birdwatcher Schema on page 2. You are asked to design a leader board for each species of Bird. The leader board ranks Observers by the number of Sightings for Birds of the given species. Write a query that computes the set of names of all Observers who are highest ranked on at least one leader board. Assume that there is no tied rankings.

جدول 1: الأسئلة المستخدمة في مجموعة بيانات UB Exam

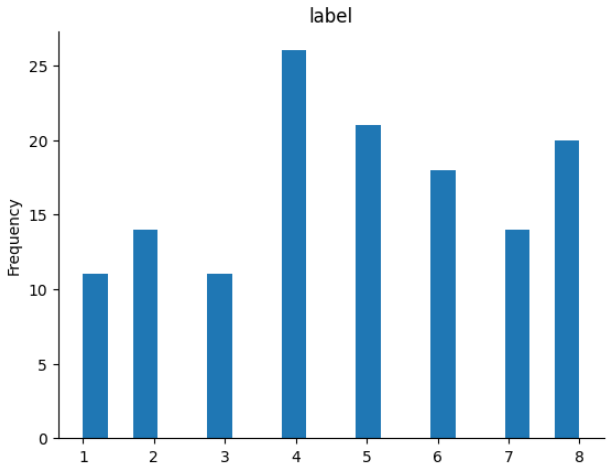
Year	2014	2015
Total number of queries	117	60
Number of syntactically correct queries	110	51
Number of distinct query strings	110	51
Number of queries with score > 50%	62	40

جدول 2: ملخص البيانات الموجودة في مجموعة بيانات UB Exam

مجموعة بياناتgoogle plus :

تم تصنيف هذه السجلات في ثماني فئات مختلفة: الحساب، النشاط، التحليلات، جهات الاتصال، الأخبار، الصيانة، الوسائط، والصور.

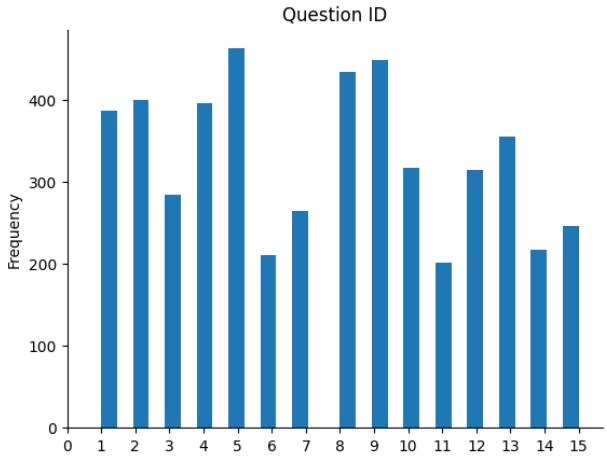
توزع الاستعلامات ضمن هذه الفئات:



الشكل 6: توزع الاستعلامات بين الفئات في مجموعة بيانات google plus

مجموعة بيانات Automated-Grading-of-SQL-Statements:

توزع الاستعلامات ضمن الفئات:



الشكل 7: توزع الاستعلامات بين الفئات في مجموعة بيانات SQL-Grading

الضبط الدقيق للنموذج:

في حالة BERT، يحتاج النموذج المدرب مسبقاً على مهام واسعة وحجم كبير من البيانات، إلى عملية ضبط دقيق لكي يعطي نتائج جيدة عند نقله إلى مهام أخرى. هذا الأسلوب يكون مفيداً خاصةً عندما تكون مجموعة البيانات محدودة للتدريب على المهمة المحددة.

قمنا باستخدام sentence-transformers وهو إطار عمل برمجي يستخدم بشكل أساسي لإيجاد تضمينات للجمل، النصوص والصور. الإطار يعتمد على PyTorch و المحولات ويقدم مجموعة واسعة من النماذج المدربة مسبقاً على مهام متنوعة. وتم تصميمه بطريقة تجعل من السهل ضبط النماذج بما يتوافق مع المهمة المراد العمل عليها.

أي أننا سنتعامل مع ما يعرف ب Sentence-BERT (SBERT) [3] وهو تعديل على شبكة BERT باستخدام عدة تنوعات كالشبكات السيامية Siamese. يتيح ذلك لـ BERT أن يستخدم في بعض المهام التي لم تكن قابلة للتطبيق سابقاً مثل مقارنة الشبه الدلالي

على نطاق واسع والتجميع واسترجاع المعلومات عبر البحث الدلالي ويتم استخدام أساليب التجميع (استخدام طبقة تجميع "mean, max pooling " لتجميع تضمين جميع الرموز) للحصول على التضمينات على مستوى الجمل وبالتالي اكتشاف التشابهات بين الجمل المدخلة.

باستخدام SBERT يتم ضبط نموذج BERT المدرب مسبقاً على مهمة تشابه الجمل زوجياً حيث أنه في هذا الشبكة، يحصل النموذج على جملتين كإدخال ويتنبأ ما إذا كانتا متشابهتين ام لا.

وبهذا فإن عملية الضبط الدقيق لا تحتاج أي تعديل على بنية نموذج BERT حيث يقوم نموذج SBERT بمهمة التعديل هذه مع وجود شبكتين كل منهما تحوي نموذج BERT مع طبقة تجميع مهمتها الحصول على تضمينات ذات طول ثابت على مستوى الجمل(الاستعلامات) ويتم استخدام هذا النموذجٍ بشكل خاص في عمليات ضبط نموذج BERT لاسيما في المهام التي تهدف إلى ايجاد التشابه بين الاستعلامات.

تحوي مجموعات البيانات التي نستخدمها بشكل أساسي الاستعلامات مع الفئة التي ينتمي اليها الاستعلام. وكوننا نتعامل مع استعلامات (لغة مهيكلة) فلا نحتاج معالجة مسبقة للبيانات بشكل موسع لاسيما أن مجموعات البيانات هي في الأساس مجموعات منقحة يدوياً من قبل الباحثين للتأكد من صحة الاستعلامات، لتحضير البيانات قمنا بما يلي:

1. معالجة مسبقة بسيطة:

- تحويل جميع المحارف إلى محارف صغيرة.
- إزالة المسافات المزدوجة.

بالنسبة لمجموعة البيانات Automated-Grading-of-SQL-Statements فقد احتاجت خطوة معالجة إضافية حيث قمنا بالاعتماد على الحقل is_correct الذي يحوي قيمة بوليانية تعبر عن كون الاستعلام الحالي صحيح كإجابة للسؤال المطروح أم لا فقمنا باسترجاع الاجابات الصحيحة فقط (الاستعلامات التي تكون صحيحة كإجابة للسؤال) مع

الاعتماد على حقل exercise_id الذي يعبر عن رقم السؤال كعنونة label وبالتالي أصبحت هذه المجموعة مماثلة لبقية المجموعات من حيث الأعمدة وتحوي عمودين query وهو الاستعلام وعمود label وهو رقم السؤال الذي تمت الاجابة عليه من خلال الاستعلام query.

2. إنشاء أزواج البيانات:

تم إنشاء مجموعة تدريب جديدة مؤقتة من مجموعات التدريب الأساسية تحوي أزواج من الاستعلامات مع قيمة ثنائية تحدد فيما اذا كان الاستعلامين من ذات الصنف(1) أم لا (0). حيث من أجل كل استعلام في مجموعة البيانات، يتم إنشاء عدة صفوف كل صف جديد يحتوي على الاستعلام نفسه واستعلام آخر من نفس التصنيف مع قيمة "1" (تشير إلى أن الاستعلامين متشابهين) حيث يتم المرور على جميع الاستعلامات من نفس الصنف مع مراعاة عدم تكرار الأزواج. ثم يتم إضافة عدد متساوٍ من الاستعلامات من تصنيفات أخرى مع قيمة "0" لكل زوج (تشير إلى أن الاستعلامين مختلفين).

قمنا بضبط نموذج Bert باستخدام أسلوب تشابه الجمل زوجياً وباستخدام الإطار المذكور، مع استخدام مقياس الخسارة التباينية.

مقياس الخسارة التباينية contrastive loss: هو نوع من دوال الخسارة المستخدمة بشكل شائع في الشبكات السيامية (المزدوجة) ونماذج أخرى لتعلم التضمينات حيث يكون الهدف هو تعلم التمثيلات التي تكون متشابهة للمدخلات المتشابهة ومختلفة للمدخلات المختلفة. في سياق تضمين الجمل أو تعلم التشابهات، يتم استخدام الخسارة التباينية لتشجيع النموذج على رسم أزواج مماثلة من الجمل بجوار بعضها البعض في الفضاء التضميني وبعيدة عن بعضها البعض للأزواج المختلفة.

تعمل دالة الخسارة التباينية من خلال مقارنة المسافات بين تضمينات أزواج المدخلات وتطبيق عقوبة استنادًا إلى مدى تشابه المدخلات أو اختلافها. فبعد الحصول على التضمينات يتم:

- حساب المسافة: لكل زوج من التضمينات، يتم حساب المقياس البعدي (مثل المسافة الإقليدية أو مسافة التشابه الجيبية cosine distance) لتقدير تشابه التضمينات.
- حساب الخسارة: بناءً على المسافة بين التضمينات وعلاماتها (مماثلة 1 أو مختلفة 0)، يتم حساب الخسارة التباينية.

بشكل رياضي، تُعرف دالة الخسارة التباينية على النحو التالي:

$$L(y,d) = (1-y) \times \frac{1}{2}\, d^2 + y \times \frac{1}{2}(\max(0, m-d))^2$$

حيث:

Y: هي التسمية التي تُشير إلى ما إذا كان الزوج متشابهاً (1) أم مختلفاً (0).

D: هو المسافة (أو درجة الاختلاف) بين تضمينات الزوج.

M: هو معامل التباعد الذي يتحكم في مدى تباعد الأزواج المختلفة، العينات السلبية يجب أن تكون لها مسافة على الأقل تساوي قيمة معامل التباعد.

تابع الخسارة المذكور يقيس الخسارة من أجل زوج واحد من الجمل ومن أجل حساب الخسارة بشكل كامل لجميع الأزواج يتم أخذ معدل الخسارة من أجل الأزواج جميعهم.

تم الاعتماد في الدراسة على القيم الإفتراضية في إطار العمل المستخدم بالنسبة لتابع المسافة وهو مسافة التشابه الجيبية.

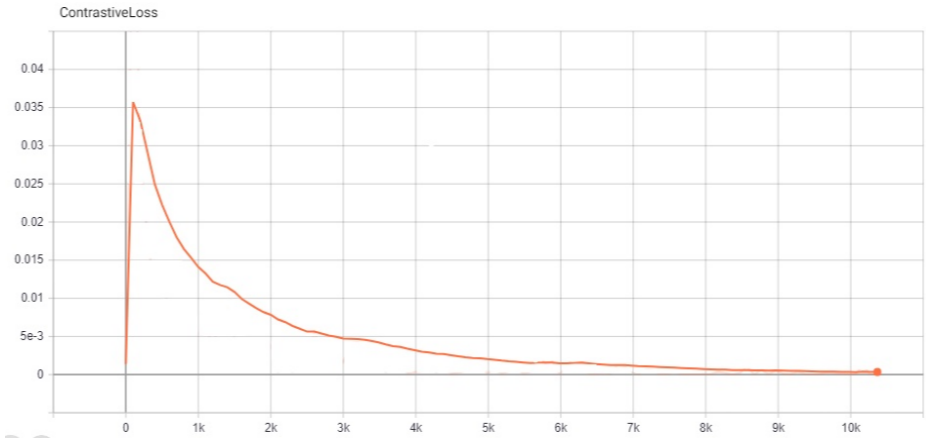
يوضح الجدول إعدادات التجربة التي قمنا بها لضبط نموذج Bert.

Epoch	3
Batch size	**2**,4
Optimizer	Adam
Learning rate	2e-5

Margin (معامل التباعد)	0.3

جدول 3: إعدادات ضبط نموذج BERT

تقييم النموذج باستخدام تابع الخسارة التباينية:



الشكل 8: الخسارة التباينة لنموذج BERT

7. مقاييس التجميع المستخدمة:

نرغب في فهم مدى قدرة تضمينات Bert على تمثيل الاستعلامات التي تقوم بنفس المهمة تمثيلا متقارباً وحتى إذا كانت مكتوبة بشكل مختلف حيث أنه ومن المعروف أنه يمكن كتابة الاستعلام بطرق مختلفة لأداء مهام متشابهة فعلى سبيل المثال الاستعلامين التاليين:

الاستعلام الأول:

```
SELECT category, SUM(sales_amount) AS total_sales
FROM sales
WHERE YEAR(sale_date) = 2023
GROUP BY category;
```

الإستعلام الثاني:

```
SELECT category, total_sales
FROM (SELECT category, SUM(sales_amount) AS total_sales
      FROM sales
      WHERE YEAR(sale_date) = 2023
      GROUP BY category) AS sales_summary;
```

الاستعلامين الأول والثاني يعتبران متشابهين حيث يتم استخدامهما لأداء ذات المهمة بالرغم من اختلاف طريقة صياغة كل منهما.

للقيام بالتقييم نقوم باستخدام مجموعة من المقاييس المستخدمة في مجال عمليات التجميع ولكننا نقوم بتقييم كل مقياس وفقاً للتسميات الحقيقية للتجميع بدلاً من تقييم نتائج التجميع نفسها أي دون استخدام خوارزمية تجميع معينة، نقوم بتقييم خطوة وسيطة وهي المسافات الزوجية(pairwise distance) لمجموعة من الاستعلامات حيث أنه باستخدام هذه المسافات ومجموعات البيانات المعنونة التي ذكرناها، يمكننا استخدام مقاييس تقييم التجميع المختلفة دون تطبيق خوارزمية معينة وذلك لفهم مدى فعالية التضمينات في توصيف مجموعة من الاستعلامات.

أي أنه يلزمنا أيضاً تابع مسافة لحساب المسافة بين التضمينات بعد تطبيق النموذج وقد اعتمدنا في دراستنا على مسافة التشابه الجيبية(cosine distance) كونها ومن خلال التجارب أدت إلى نتائج أفضل من التوابع الأخرى كالمسافة الإقليدية، مانهاتن وغيرها.

تستخدم مقاييس تقييم التجميع بشكل خاص للتحقق من جودة المجموعات الناتجة عن طريق تقدير درجة تماسك العينات التي تنتمي إلى نفس المجموعة إضافة إلى تقدير درجة الفصل بين العينات التي تنتمي إلى مجموعات مختلفة. ونتيجة لذلك، سنستخدم ثلاث

مقاييس للتحقق من التجميع بما في ذلك معامل السيلويت المتوسط (Silhouette Coefficient)، ومؤشر Betacv، ومؤشر دان (Dunn index)، حيث تقيم كل منها الصفات المذكورة أعلاه في صياغتها. ونقوم بمقارنة النتائج مع أساليب التشابه الثلاث (Aligon، Aouiche، Makiyama).

كما نقوم باستخدام هذه المقاييس لتقييم التجميعات الناتجة عن نموذج BERT وعدد من النماذج اللغوية الأخرى اعتمادا على ثلاث خوارزميات تجميع Clustering هي Kmeans, HAC, Optics.

مؤشر Dunn:

هو مقياس يستخدم لتقدير جودة التجميع، إذا كانت القيمة عالية فإن هذا يشير إلى أن العناصر داخل المجموعة تمتاز بشكل كبير عن العناصر في المجموعات الأخرى، مما يشير إلى جودة تجميع جيدة. على العكس من ذلك، إذا كانت القيمة منخفضة. يعطى المؤشر بالعلاقة:

$$\text{Dunn Index} = \min(\text{inter-cluster distances}) / \max(\text{intra-cluster distances})$$

حيث أن:

Inter-cluster distances: مصطلح يشير إلى المسافة بين مجموعات مختلفة. ويقيس الفارق أو عدم التشابه بين المجموعات. كلما زادت المسافة بين المجموعات، زاد تمييزها وانفصالها عن بعضها البعض. عادةً ما يتم حساب المسافة بين المراكز (النقاط الوسطى أو المراكز) للمجموعات أو كأدنى مسافة بين نقاط البيانات في مجموعات مختلفة.

Intra-cluster distances: هذا المصطلح يشير إلى المسافة بين نقاط البيانات داخل نفس المجموعة. بمعنى آخر، يقيس التماسك أو الانسجام بين نقاط البيانات داخل مجموعة واحدة. كلما كانت المسافة داخل المجموعة أقل، زاد تشابه وتجمع نقاط البيانات داخل

المجموعة. يتم حساب المسافة داخل المجموعة عادةً كالمسافة المتوسطة أو القصوى بين جميع أزواج نقاط البيانات داخل المجموعة.

معامل Silhouette Coefficient:

معامل الظل (Silhouette Coefficient) هو مقياس يستخدم لتقدير جودة التجميع، يوفر هذا المعامل تقييماً لمدى تماسك المجموعات. لحساب معامل الظل، يتم قياس مدى تشابه نقاط المجموعة الفردية مقارنة بنقاط المجموعات الأخرى. يكون معامل الظل في النطاق من −1 إلى 1. القيم الإيجابية تشير إلى أن النقاط داخل المجموعة أقرب إلى بعضها البعض من متوسط المسافة بين المجموعات وكلما كانت القيمة قريبة من 1، فهذا يشير إلى أن العنصر موجود في المجموعة المناسبة وأنهُ قريب جداً من العناصر الأخرى في نفس المجموعة، في حين أن القيم السالبة تشير إلى أن هناك تشابهاً أقل بين نقاط المجموعة مما هو متوسط المسافة بين المجموعات.

في مجموعة ما، يمكن حساب معامل الظل للنقطة i بالشكل التالي:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

حيث:

- S(i) هو معامل الظل للنقطة i.
- a(i) هو المسافة المتوسطة بين نقطة i ونقاط المجموعة التي تحتوي عليها.
- b(i) هو المسافة المتوسطة بين نقطة i وأقرب مجموعة تحتوي على نقاط لا تنتمي إليها.

يمكن حساب معامل الظل لكل النقاط في المجموعة، ومن ثم يتم حساب المعامل الإجمالي بواسطة القيمة المتوسطة لمعاملات الظل لجميع النقاط.

مقياس Betacv:

يستخدم لتقدير جودة التجميع، يقيس مدى الجودة بناءً على ارتباطها المحموعة (intra-distance) واستقلالها (inter-distance). يعتبر Beta_cv هو النسبة بين متوسط المسافة داخل المجموعة ومتوسط المسافة بين المجموعات. القيم الصغيرة للمقياس تشير إلى جودة أفضل لعملية التجميع.

8. التجربة والنتائج:

تم إجراء التجربة وتنفيذها على منصة Google Colaboratory السحابية واستخدام sentence_transformers ومكتبات sklearn، beta_cv لاستخدام مقاييس الأداء منهما.

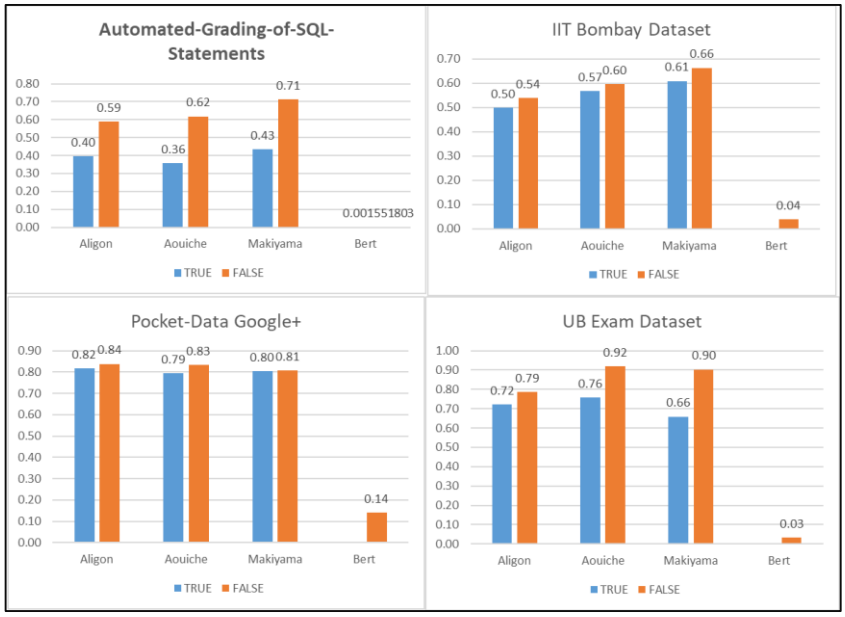
التجربة الأولى:

الهدف معرفة كفاءة تضمينات نموذج Bert بالنسبة لايجاد التشابه بين الاستعلامات مقارنة بالاساليب المذكورة في الدراسات السابقة. بدايةً قمنا بتمرير الاستعلامات الى النموذج المدرب وحصلنا على التضمينات الخاصة بهذه الاستعلامات. في الخطوة التالية قمنا بتمرير مجموعة التضمينات التي حصلنا عليها مع عنوان المجموعة التي من المفترض أن ينتمي إليها كل استعلام (كوننا نتعامل مع مجموعات بيانات معنونة) إضافة الى تحديد تابع المسافة (جيب التمام cosine distance) إلى مقاييس الأداء المعرفة ضمن المكتبات (betacv، silhouette، Dunn index) لتقييم أداء النموذج.

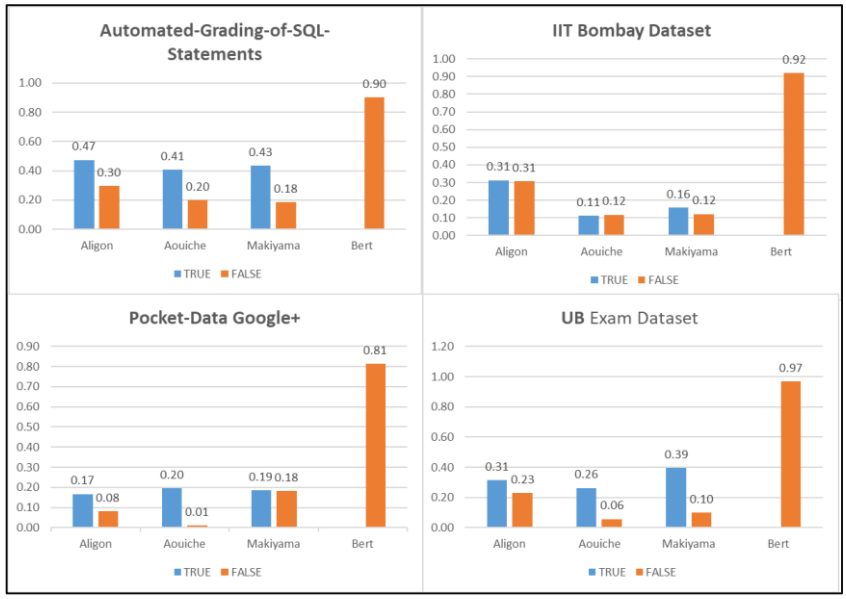
تظهر الأشكال التالية مقارنة بين ثلاثة أساليب تشابه مذكورة في المرجعيات (Aligon، Aouiche، Makiyama) كل منها يستخدم أسلوب لتمثيل الاستعلام مع تابع مسافة

معين لقياس التشابه، والتشابه بين الاستعلامات باستخدام تضمينات Bert، حيث أن الأعمدة التي تعبر عن القيمة False هي دون تطبيق خطوة المعالجة المسبقة للإستعلامات (خطوة استخراج الميزات) المقترحة في الدراسة [2] بينما القيمة True تعبر عن إجراء هذه المعالجة قبل تطبيق مقياس التشابه والجدير بالذكر أننا نستخدم نموذج Bert الذي قمنا بضبطه دون تطبيق أي معالجة مسبقة للإستعلامات وذلك لأن هدف البحث هو تسليط الضوء على مدى فعالية التضمينات المستخرجة من نموذج Bert في تمثيل الاستعلامات دون أي تدخل بخطوات معالجة إضافية للبيانات مصممة يدوياً.

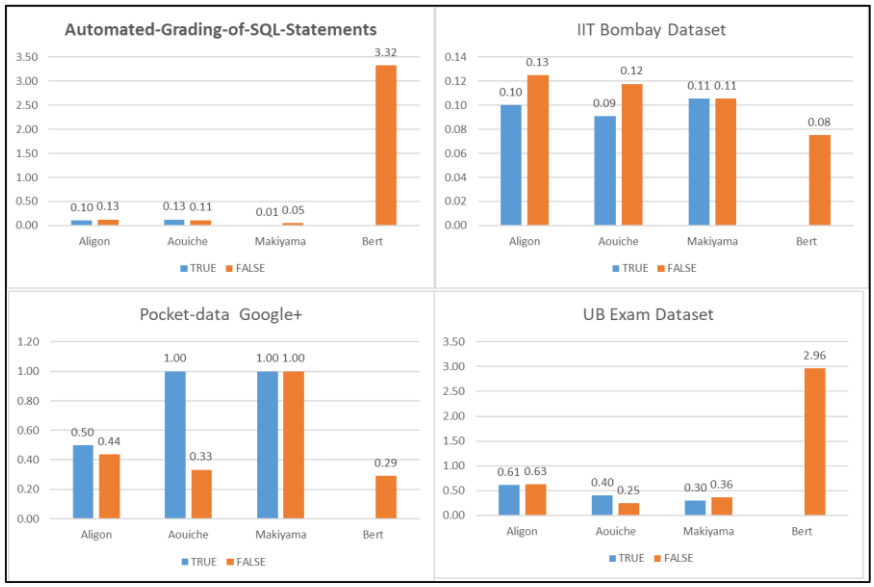
نتائج التجربة الأولى:



الشكل 9: الأداء حسب مقياس beta_cv (القيم الأصغر هي الأفضل)



الشكل 10: متوسط silhouette coefficient(القيم الأكبر هي الأفضل)



الشكل 11: الأداء حسب Dunn index (القيم الأكبر هي الأفضل)

مناقشة نتائج التجربة الأولى:

أظهرت النتائج التفوق الواضح لنموذج Bert على الأساليب الأخرى (Aligon، Aouiche، Makiyama) حيث حصل نموذج Bert على أعلى قيم لمعامل Silhouette وصلت لتحسين لا يقل عن 90% كحد أدنى (في مجموعة بيانات automated-grading-sql-statements). كما وتفوق بالنسبة لمقياس Betacv عبر جميع مجموعات البيانات الأربع المختبرة . بنسبة لا تقل عن 82%. مما يشير إلى قدرة النموذج الفائقة في تمييز الاستعلامات المتشابهة وغير المتشابهة بشكل فعال. تفوقه يعكس قدرة النموذج على تكوين تجميعات متماسكة ومنفصلة بوضوح عن بعضها البعض. بالمقابل، أظهرت الأساليب الأخرى أداءً أقل وتباينت في اعتمادها على خطوات المعالجة المسبقة لتحسين النتائج، مما يوضح محدودية هذه الأساليب مقارنة بـ Bert.

أما بالنسبة ل Dunn index تفوق نموذج Bert في مجموعتين من بيانات الاختبار حيث أظهر قدرة عالية في تكوين تجميعات ذات جودة أفضل. ومع ذلك، كان أداءه أقل مقارنة بالنماذج التقليدية في مجموعات البيانات الأخرى مثل Pocket-Data Google+ و UB Exam Dataset تعكس هذه النتائج أن نموذج Bert يمكن أن يكون فعالًا جدًا في بعض الحالات، بينما قد تكون النماذج التقليدية أكثر فعالية في حالات أخرى، مما يشير إلى أن اختيار النموذج المناسب يعتمد بشكل كبير على طبيعة البيانات المستخدمة.

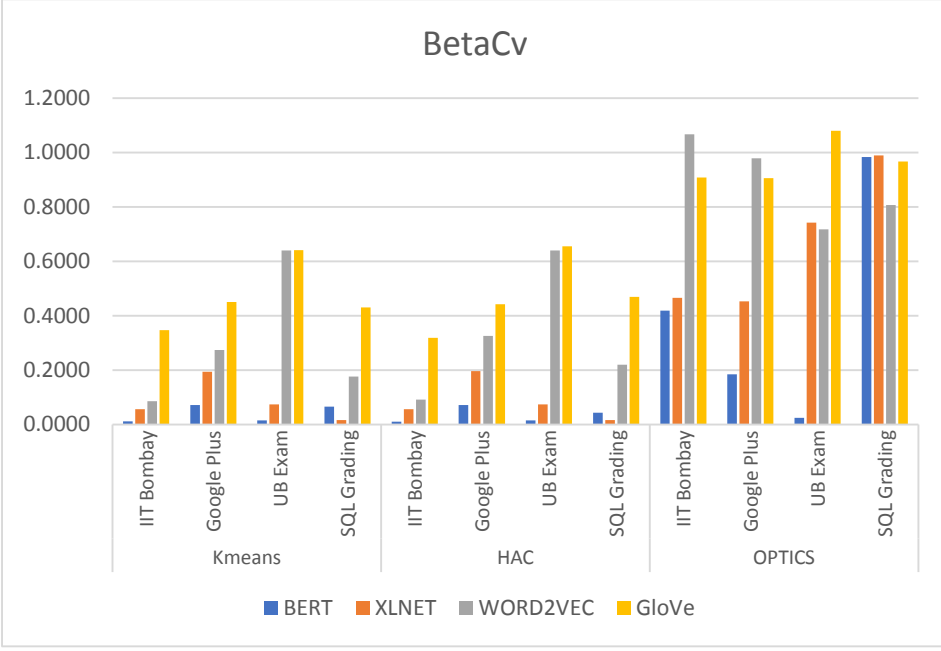
التجربة الثانية:

في هذه التجربة، قمنا بتقييم أداء نماذج التضمين في سياق التجميع باستخدام ثلاثة خوارزميات تجميع مختلفة Optics, Kmeans و HAC. ركزنا بشكل خاص على Bert ونماذج أخرى مثل XLNet وWord2Vec و GloVeوتم تطبيق كل نموذج من النماذج الأربعة على مجموعات البيانات الأربعة المستخدمة في البحث. كان الهدف من التجربة هو تحديد النموذج الذي يقدم أفضل تمثيلات نصية تساعد في تشكيل تجمعات واضحة ومتمايزة.

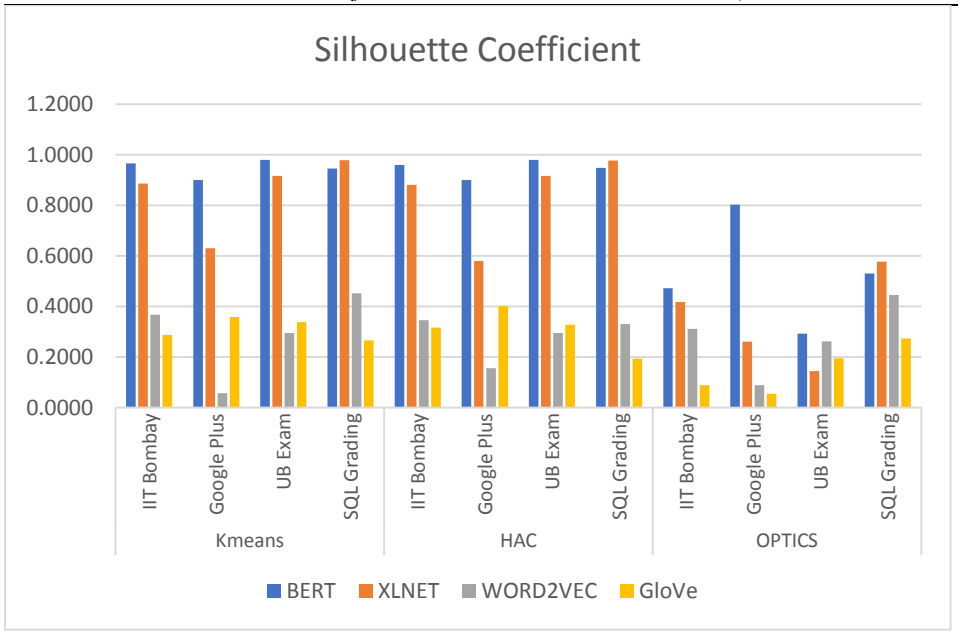
حيث يتم بداية استخراج التضمينات من النموذج المختبر ومن ثم تطبيق خوارزمية التجميع (جميع الخوارزميات بالتسلسل) وفي النهاية يتم حساب مقاييس الأداء (مقاييس التجميع المعتمدة في بحثنا Dunn Index, Betacv, ومعامل Silhouette.

نتائج التجربة الثانية:

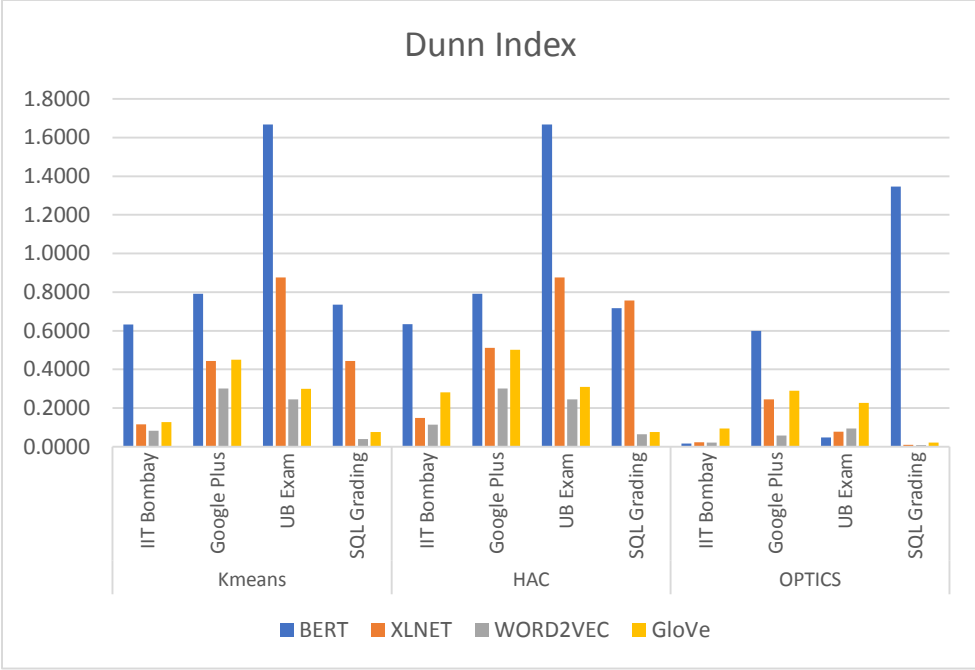
ملاحظة: في مخططات النتائج لهذه التجربة تم اختصار تسمية مجموعة البيانات automated-grading-sql-statements إلى SQL Grading بغرض التخفيف من تعقيد عرض المخططات.



الشكل 12: الأداء حسب مقياس beta_cv (القيم الأصغر هي الأفضل)



الشكل 13: متوسط silhouette coefficient (القيم الأكبر هي الأفضل)



الشكل 14: الأداء حسب Dunn index (القيم الأكبر هي الأفضل)

مناقشة نتائج التجربة الثانية:

نلاحظ انه وبالنسبة لمقياس Betacv أظهرت النتائج أن نموذج BERT هو الأكثر تفوقاً في معظم الحالات عبر جميع خوارزميات التجميع في ثلاث مجموعات اختبار مع نتائج مقاربة لأفضل النماذج في مجموعة البيانات الأخيرة Automated-Grading-SQL-Statements.

بالنسبة ل Dunn Index نلاحظ من النتائج السابقة أنBERT يتفوق بشكل ملحوظ في تشكيل تجمعات متماسكة ومنفصلة عبر مختلف مجموعات البيانات، خاصةً في خوارزميات Kmeans وHAC مع ملاحظة أداء أقل بالنسبة لخوارزمية Optics ويمكن أن يعزى السبب إلى التعقيد الحسابي حيث أن Optics أكثر تعقيداً من حيث الحسابات مقارنة مثلا بـK-means ، وقد تواجه صعوبة في التعامل مع تمثيلات البيانات العالية الأبعاد التي ينتجهاBERT .

أما بالنسبة لمعامل silhouette الذي يقيس مدى تشابه النقاط داخل التجمع الواحد مقارنةً بتشابهها مع النقاط في التجمعات الأخرى.

يتضح من النتائج أن BERT يظهر أداءً قوياً ومتسقاً عبر مختلف مجموعات البيانات وخوارزميات التجميع، باستثناء بعض الحالات لاسيما مجموعة بيانات Automated-Grading-SQL-Statements التي تفوق فيها نموذج XLNET .

9. الاستنتاجات والتوصيات:

أظهر استخدام تضمينات BERT لاستعلامات SQL أداءً جيداً حيث تمكن النموذج من توفير تمثيلات سياقية غنية أثبتت جدواها في مجال تقييم الشبه بين الاستعلامات حيث تشير النتائج بقوة إلى أنه يمكن باستخدام تضمينات BERT استكشاف العلاقات والأنماط المعقدة في استعلامات SQL.

تفوق النموذج في أداءه على أداء الأساليب الموجودة في المرجعيات حيث كان هناك تفوق ملاحظ في مقياس Betacv مما يعني أن تضمينات BERT تلتقط بشكل فعال كل من الاتساق والفصل بين التجمعات في البيانات حيث يعتبر وجود تجمعات جيدة الفصل ومتميزة ضرورياً للقيام بتحليلات دقيقة. كما تفوق على عدد من النماذج اللغوية فيما يتعلق بالتجميع Clustering عبر عدة خوارزمية وأظهر النتائج نموذج XLNET كمنافس قوي لنموذج BERT.

علاوة على ذلك، يؤكد التفوق الملاحظ لنموذج BERT في مقياس silhouette على قدرته على إنتاج تجمعات متماسكة داخلياً بشكل أفضل مقارنة بالأساليب السابقة والنماذج اللغوية الأخرى. تعتبر هذه النتيجة مهمة لفهم الهيكل الأساسي والأنماط ضمن استعلامات SQL.

التحسين الكبير المقدم من النموذج يشير إلى أن التضمينات التي يقدمها نموذج BERT تضيف مستوى جديداً لفهم الاستعلامات المطبقة واستخراج الأنماط ويمكن أن يرجع النجاح

إلى كفاءة النموذج في توسيع السياق وفهم دلالات الاستعلامات، متجاوزاً بذلك قدرات الأساليب المعتمدة على النحو.

10.الأعمال المستقبلية:

يمكن توسيع الدراسة لتشمل إجراء مقارنة شاملة بين تضمينات BERT ونماذج لغوية أخرى. كما يمكن تقييم فعالية هذه التضمينات في مهام أخرى مثل مهام ضغط الأحمال التي تعتمد في جوهرها على تجميع Clustering الاستعلامات واستخراج عينة معبرة من الحمل (مجموعة من الاستعلامات المطبقة على القاعدة). إضافة إلى إمكانية تقييم أداء النموذج الذي تم ضبطه في هذا البحث في مهام إدارة قواعد البيانات كمهمة ضبط الفهارس.

11.المراجع:

[1] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. 2020- Pre-trained models for natural language processing: A survey. **Science China Technological Sciences,** Vol.63, 10, 1872–1897.

[2] Kul G., Luong, D. T. A., Xie, T., Chandola, V., Kennedy, O., Upadhyaya, S. 2018-Similarity Metrics for SQL Query Clustering. **IEEE Transactions on Knowledge and Data Engineering**, Vol.*30*, 12, 2408–2420.

[3] Reimers, N., Gurevych, I. 2019- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. **arXiv (Cornell University).**

[4] MacQueen, J. B.1967- Some methods for classification and analysis of multivariate observations. **In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability**, Vol.1, 14, 281–297.

[5] Ankerst, M., Breunig, M., Kriegel, H., Sander, J. 1999- Optics: Ordering points to identify the clustering structure, **ACM Sigmod record**, Vol.28, 2, 49–60.

[6] Lukasová, A. 1979- Hierarchical agglomerative clustering procedure. **Pattern Recognition**, Vol.11, 5–6, 365–381.

[7] Devlin, J., Chang, M., Lee, K., Toutanova, K. 2018- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv (Cornell University).**

[8] Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S. 2015- Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. **arXiv (Cornell University),** 19–27.

[9] Trummer, I. 2022- DB-BERT: A Database Tuning Tool that Reads the Manual. **Proceedings of the 2022 International Conference on Management of Data**.

[10] Hagglund, M., Pena, F. J. R., Pashami, S., Al-Shishtawy, A., Payberah, A. H. 2021- COCLUBERT: Clustering Machine Learning Source Code. **2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA).**

[11] Chandra, B., Joseph, M., Radhakrishnan, B., Acharya, S., Sudarshan, S. 2016- Partial marking for automated grading of SQL queries. **Proceedings of the VLDB Endowment**, Vol.9, 13, 1541–1544.

[12] Jain, S., Howe, B. 2018- Query2VEC: NLP meets databases for generalized workload analytics. **arXiv (Cornell University).**

[13] Aligon, J., Golfarelli, M., Marcel, P., Rizzi, S., Turricchia, E. 2013- Similarity measures for OLAP sessions. **Knowledge and Information Systems,** Vol.39, 2, 463–489.

[14] Tahmasebi, S., Payberah, A. H., Soylu, A., Roman, D., Matskin, M. 2022- TranSQL: A Transformer-based Model for Classifying SQL Queries. **2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)**.

[15] Aouiche, K., Jouve, P., & Darmont, J. 2006- Clustering-Based materialized view selection in data warehouses**. In Lecture Notes in Computer Science**, 81–95.

[16] V. H. Makiyama, M. J. Raddick, and R. D. Santos, 2015- Text mining applied to SQL queries: A case study for the SDSS SkyServer, **in SIMBig**.

[17] B. Chandra, B. Chawda, B. Kar, K. V. Reddy, S. Shah, and S. Sudarshan, “Data generation for testing and grading SQL queries,” VLDBj, 2015.

[18] Wang, J.: Dataset: combining dynamic and static analysis for automated grading sql statements (2020).URL https://doi.org/10.5281/zenodo.6526769

[19] J. MacQueen et al., Some methods for classification and analysis of multivariate observations, **in Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[20] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, Optics: Ordering points to identify the clustering structure, **ACM Sigmod record**, vol. 28, no. 2, pp. 49–60, 1999

[21] A. Lukasova, Hierarchical agglomerative clustering procedure, Pattern Recognition, vol. 11, no. 5-6, pp. 365–381, 1979.

[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," arXiv preprint arXiv:1310.4546, 2013.

[23] J. Pennington, R. Socher, and C. D. Manning, 2014 - Glove: Global vectors for word representation, in **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),** 2014, pp. 1532–1543.

[24] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V., 2019- XLNet: Generalized autoregressive pretraining for language understanding. **arXiv preprint arXiv:1906.08237**.

[25] Brian A. Cumi-Guzman, Alejandro D. Espinosa-Chim, Mauricio G. Orozco-del-Castillo, and Juan A. Recio-García, 2024-Counterfactual Explanation of a Classification Model for Detecting SQL Injection Attacks. **ICCBR'24 Workshop Proceedings.**