**جامعة تشرين كلية الهندسة المعلوماتية قسم الذكاء الصنعي**

الجمهورية العربية السورية

وزارة التعليم العالي

جامعة تشرين \_ كلية الهندسة المعلوماتية

قسم الذكاء الصنعي

**الترجمة من اللغة الإنكليزية إلى اللغة الفرنسية**

**Translation from English to French**

( مشروع فصلي )

**إعداد الطلاب :**

عمار معلا زين العابدين سلمان روان دريباتي

**إشراف :**

العام الدراسي 2023-2024

**ملخص**

في هذا المشروع، تم استخدام تقنيات التعلم الآلي لتحسين أداء الترجمة من اللغة الإنجليزية إلى اللغة الفرنسية، وتم مقارنة أداء شبكات النموذج العصبي العميق التكراري (RNN) مع وبدون تضمين الكلمات في المدخلات (embedding).

أظهرت النتائج أن استخدام تضمين الكلمات في المدخلات (embedding) يحسن أداء النماذج في الترجمة بشكل كبير، حيث تمكنت النماذج التي استخدمت التضمينات من تحقيق معدل دقة أفضل بكثير مقارنة بالنماذج التي لم تستخدم التضمينات.

**Abstract**

In this project, machine learning techniques were used to improve the performance of English to French translation, and the performance of Recurrent Neural Networks (RNNs) with and without word embedding in the input was compared.

The results showed that using word embedding in the input significantly improves the performance of the models in translation. Models that used embeddings were able to achieve much higher accuracy rates compared to models that did not use embeddings.

**الفهرس**

**1-الفصل الأول :** مقدمة..............................................................................1

1-1-الدراسات المرجعية..........................................................................1

1-1-1-الأوراق التي استندنا عليها خلال المشروع:............................................1

1-1-2- جدول التلخيص : ..................................................................4

2-1- البيئة والأدوات المستخدمة :................................................................5

3-1- منهجية البحث :...........................................................................6

**2-الفصل الثاني :** الجزء النظري :....................................................................7

2-1-خوارزميات الترجمة ما قبل 2014 :..........................................................8

2-1-1- Knowledge-based algorithms:.................................................9

2-1-2- Statistical Machine Translation :............................................10

2-2- خوارزميات الترجمة ما بعد 2014 : ........................................................11

2-2-1- Neural-based : .................................................................................11

2-2-2- Deep Learning-based Neural Machine Translation:.......................12

2-2-3- Artificial Intelligence-based Models :.......................................13

**3-الفصل الثالث :** الجزء العملي :..................................................................14

3-1-

3-1-1-

3-1-2-

3-2-النتائج:..................................................................................25

3-3-الصعوبات:..............................................................................29

3-4-التوصيات:...............................................................................30

**المراجع Refrences ..............................................................................**31

**فهرس الأشكال :**

الشكل 3-1 .......................................................................................12

الشكل 3-2 .......................................................................................13

الشكل 3-3 .......................................................................................14

الشكل 3-4 .......................................................................................16

الشكل 3-5 .......................................................................................16

الشكل 3-6 .......................................................................................17

الشكل 3-7 .......................................................................................17

الشكل 3-8 .......................................................................................13

الشكل 3-9 .......................................................................................20

الشكل 3-10 .....................................................................................24

**فهرس الجداول :**

الجدول(1-1) ........................................................................................5

الجدول(2-1) ......................................................................................12

**قائمة المصطلحات:**

|  |  |
| --- | --- |
| **المختصر الموافق** | **المصطلح** |
| RNN | Recurrent Neural Networks |
| DNN | Deep Neural Networks |
| LSTM | Long Short-Term Memory |
| MLOps | Machine Learning Operations |
|  |  |
|  |  |
|  |  |

1. **الفصل الأول: مقدمة**

الترجمة من لغة إلى أخرى تعتبر واحدة من أهم مهام وتحديات تعلم الآلة ومعالجة اللغات الطبيعية. فاللغات الطبيعية تتضمن تنوعًا كبيرًا في النحو والمفردات والتعابير اللغوية، وهذا يجعل من الترجمة بين اللغات المختلفة مهمة صعبة.

تعتمد تقنيات الترجمة الحديثة على النماذج العميقة للتعلم الآلي، والتي تتضمن شبكات النموذج العصبي العميق (Deep Neural Networks) وخوارزميات تعلم الآلة المعتمدة على البيانات. وتستخدم هذه التقنيات مجموعات كبيرة من البيانات المترجمة لتدريب نماذج الترجمة الآلية، حيث يتم تدريب النموذج على تحويل النص المكتوب في اللغة الأولى (الإنجليزية) إلى النص المكتوب في اللغة الثانية (الفرنسية).

تعتمد نماذج الترجمة الحديثة على تقنيات مثل تضمين الكلمات (Word Embedding) لتحقيق أداء ممتاز في الترجمة الآلية. ومع تزايد حجم مجموعات البيانات المترجمة وتحسن تقنيات التعلم الآلي، يمكن توفير نتائج الترجمة الآلية التي تقارب جودة الترجمة البشرية في بعض الحالات، وهذا يجعل الترجمة الآلية أداة قيمة في مجالات الأعمال والتعليم والتواصل الدولي.

* 1. **الدراسات المرجعية :**

**1-1-1-الأوراق التي استندنا عليها خلال المشروع:**

Efficient Estimation of Word Representations in Vector Space-

في هذه الورقة تم انتاج طريقة عصرية لتمثيل النص ضمن vector space structured هيكليته تعتمد

على العلاقات الدلالية بين الكلمات أو بشكل أكثر وضوح كل كلمة سيتم ربطها بشعاع ثم إسقاطها في هذا

الفضاء بحيث تكون العلاقات الهندسية بين word vectors انعكاس للعلاقات الدلالية بين الكلمات, وتعتبر

هذه التقنية أساسية في تطوير نماذج الترجمة الألية باستخدام الشبكات العصبونية, حيث يتم استخدام

word embedding لتمثيل الجمل في شكل رقمي بحيث يمكن للشبكات العصبية فهم المعنى والمضمون الخاص بالجمل والكلمات.

تم تأليف الورقة من قبل Thomas Mikolov, Quoc Le, Edward H ، وتم نشرها في عام 2013 في مؤتمر تقنيات اللغة الطبيعية.

- Sequence to Sequence Learning with Neural Networks

في هذه الورقة تم استخدام الشبكات العصبونية LSTM للترجمة من جملة إلى أخرى معتمدين على نموذج

Sequence-to-Sequence حيث يتم في البداية استخدام word embedding لتمثيل الجمل ثم يتم

إدخال هذه المتجهات إلى شبكة عصبونية والتي تقوم بتحويل هذه المتجهات إلى المتجهات الخاصة بالجملة

الهدف وبعد ذلك يتم تحويل المتجهات الناتجة إلى كلمات حقيقية في اللغة المستهدفة وتم الاعتماد أيضا

على طبقة Attention للتركيز على الكلمات الأكثر أهمية من خلال تخصيص وزن معين لكل كلمة في الجملة.

تم تأليف الورقة من قبل Ilya Sutskever, Oriol Vinyals,Quoc V. Le ، وتم نشرها في عام 2014.

- Neural Machine Translation by Jointly Learning to Align and Translate

في هذه الورقة تم استخدام نفس الآلية التي تم الاعتماد عليها في الورقة السابقة من استخدام تقنية التضمين و شبكة LSTM وكذلك استخدام طبقة Attention للتركيز على الكلمات الأكثر أهمية ومنعا لحدوث أخطاء في الترجمة مثل وضع كلمة ترجمه بغير موضعها الصحيح مثل لو كانت جملة الترجمة "الكتاب على الطاولة" فتكون الترجمة "The table is on the book" بدلا من "The book is on the table"

وهذا خاطئ لأنه تم وضع كلمتين بغير موضعها الصحيح لذلك تم استخدام طبقة Alignment لتحديد الموضع المناسب لكل كلمة اعتماد على درجات التركيز التي تأتي طبقة Attention.

تم تأليف الورقة من قبل Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio , وتم نشرها في عام 2014.

|  |  |  |  |
| --- | --- | --- | --- |
| **مساؤى** | **محاسن** | **الأفكار الرئيسية** | **اسم الورقة البحثية** |
| قد تعاني طريقة التضمين من بعض المشاكل في تمثيل الكلمات النادرة ويحتاج أيضا مصادر بيانات كبيرة | تقليل حجم النماذج اللازمة لتمثيل الكلمات وتحسين وقت التدريب للنماذج العصبية , مما يساعد في تحسين أداء تطبيقات الترجمة الآلية | استخدام تقنية التضمين لتحسين جودة الترجمة الآلية باستخدام الشبكات العصبونية | Efficient Estimation of Word Representations in Vector Space |
| تعقيد عملية التدريب والحاجة إلى كميات كبيرة من البيانات لضمان تحقيق النتائج المرجوة وظهور الأخطاء في الترجمة الآلية والتي تؤثر على جودة الترجمة وتفاصيل المعنى | القدرة على التعامل مع مشكلة طول الجمل والتعامل مع ترجمة الجمل ذات المعاني المجردة والمعقدة بفضل طبقة attention | استخدام الشبكة العصبونية LSTM مع تقنية التضمين وتقنية attention للتركيز على الكلمات الأكثر أهمية بالنسبة للجملة المستهدفة | Sequence to Sequence Learning with Neural Networks |
| على الرغم من استخدام Alignmentإلا أن النموذج يعاني من مشكلة في التعامل مع الجمل المعقدة والطويلة وخاصة الجمل التي تحوي كلمات متعددة المعاني حيث تؤدي التداخلات بين الكلمات إلى تحديد مواضع غير ملائمة للكلمات | تحسين جودة ودقة الترجمة والتغلب على مشكلة الترجمة الحرفية بعد استخدام تقنية Alignment | استخدام الشبكة العصبونية LSTM مع تقنية التضمين وتقنية attention وتقنية Alignment لتحديد الموضع المناسب للكلمات في الجملة الهدف بشكل متزامن | Neural Machine Translation by Jointly Learning to Align and Translate |

**-2-1البيئة والأدوات المستخدمة**

في هذه الفقرة، سنشرح بإيجاز عن الأدوات والبيئة التي تم استخدامها في هذا البحث، والتي تشمل

Github Actions، Docker، Python، Jupyter Notebook، FastAPI، Colab، TensorFlow، Keras، NumPy، PyYAML، Pickle.

* Github Actions هو خدمة تقدمها Github لإدارة عمليات التطوير الخاصة بمشاريع البرمجة. باستخدام Github Actions يمكنك إنشاء نظام للتحكم بعمليات التطوير والإختبار والنشر بصورة آلية، بحيث يمكنك تحسين الإنتاجية وتوفير الوقت اللازم لإنجاز المهام. يتم إعداد Github Actions بإستخدام ملفات YAML التي تحتوي على سلسلة من الخطوات التي يجب تنفيذها.
* Docker هو برنامج يساعد على إنشاء وتشغيل بيئات البرمجة في أي نظام تشغيل، بما في ذلك Linux وWindows وMacOS. يساعد Docker في توفير بيئة متكاملة للتطوير والتشغيل، ويسمح بإنشاء حاويات (containers) تحتوي على جميع المكتبات والأدوات والإعدادات اللازمة لتشغيل تطبيقك.
* Python هي لغة برمجة شائعة الاستخدام ومفتوحة المصدر، وتستخدم في العديد من المجالات بما في ذلك الذكاء الاصطناعي وعلوم البيانات وتطوير الويب.
* Jupyter Notebook هو برنامج تفاعلي للبرمجة يسمح للمستخدمين بكتابة الشفرات وعرض النتائج في واجهة مستخدم بديهية. يتم استخدام Jupyter Notebook عادة للتعلم العملي والتجارب السريعة في علوم البيانات والذكاء الاصطناعي.
* Colab هو بيئة تفاعلية تقدمها Google، ويتم استخدامها عادة لتطوير النماذج الذكية والتعلم العميق وعلوم البيانات. يتميز Colab بواجهة مستخدم بديهية وخدمات السحابة التي يتم توفيرها من قبل Google.
* TensorFlow هو إطار عمل للتعلم الآلي والذكاء الاصطناعي، ويستخدم عادة لبناء الشبكات العصبية وتدريبها.
* Keras هو إطار عمل للتعلم الآلي يُستخدم كواجهة عالية المستوى لتطبيقات TensorFlow. يعتبر Keras سهل الاستخدام ويتميز بالمرونة والسرعة في بناء الشبكات العصبية وتدريبها.
* NumPy هو مكتبة لغة بايثون مفتوحة المصدر تُستخدم للعمليات الرياضية والعلمية، وتتضمن مجموعة من الأدوات والوظائف المفيدة لتحليل ومعالجة البيانات الكبيرة. يتميز NumPy بالسرعة والكفاءة والتعامل مع البيانات في صيغة المصفوفات.
* PyYAML هو مكتبة لغة بايثون مفتوحة المصدر يتم استخدامها للتعامل مع صيغة YAML يستخدم YAML عادة لتخزين البيانات المنظمة بشكل هرمي، ويمكن استخدام PyYAML لتحويل البيانات بين صيغ YAML وصيغ أخرى.
* Pickle هو مكتبة لغة بايثون تُستخدم للتسلسل والإلغاء التسلسل للبيانات. يستخدم Pickle عادة لحفظ واستعادة البيانات في صيغة ملف، ويتميز بالسهولة والكفاءة في التعامل مع البيانات.
* FastAPI هو إطار عمل لتطوير واجهات برمجة تطبيقات RESTful بسرعة وكفاءة باستخدام لغة Python. يتميز FastAPI بأدائه العالي ووجود وثائق دقيقة وشاملة.

**-3-1 منهجية البحث**

Dataset: English to Frence

معالجة الداتا

استخدام الشبكة العصبونية

LSTM

تحويل الجملة الانكليزية إلى الفرنسية

Dataset: English to Frence

معالجة الداتا

استخدام تقنية التضمين   
Word Embedding

استخدام الشبكة العصبونية

LSTM

تحويل الجملة الانكليزية إلى الفرنسية

**1-1-2-جدول التلخيص :**

|  |  |  |  |
| --- | --- | --- | --- |
| اسم الورقة البحثية | الأفكار الرئيسية | الخوارزميات | Dataset |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

الجدول(1-1)

**2-الفصل الثاني : الجزء النظري :**

بداية سنتحدث عن الخوارزميات التي استخدمت في مشروع الترجمة من اللغة الإنكليزية إلى اللغة الفرنسية من الأقدم للأحدث , حيث قام العلماء بتطوير الخوارزميات انطلاقآ من المشاكل الظاهرة محاولة تخفيفها.

يوجد فترتان تاريخيتان للترجمة من اللغة الإنكليزية إلى اللغة الفرنسية , قبل 2014 وفترة الترجمة القائمة على التعلم العميق بعد 2014 .

في أي ترجمة، سواء كانت بشرية أو آلية، يجب أن ينقل معنى النص في لغة المصدر (الأصلية) بالكامل إلى معناه المعادل في ترجمة اللغة الهدف. وبينما يبدو هذا واضحا ظاهريا، فهو غالبا ما يكون أكثر تعقيدا. فالترجمة ليست مجرد استبدال كلمة بكلمة.

يجب على المترجم البشري تفسير وتحليل جميع العناصر داخل النص وفهم كيف يمكن لكل كلمة أن تؤثر على سياق النص. وهذا يتطلب خبرة واسعة في قواعد اللغة، وبناء الجمل (تراكيب الجمل)، والدلالات (المعاني)، وما إلى ذلك، في اللغتين، المصدر والهدف، بالاضافة إلى الخبرة في مجال النص.

كما أن لكل من الترجمة البشرية والآلية حصتهما من التحديات. على سبيل المثال، لن ينتج اثنين من المترجمين ترجمات متطابقة للنص نفسه في نفس اللغتين، كما أن الأمر قد يستغرق جولات تنقيح عديدة لتلبية متطلبات العميل. أما الترجمات الآلية فتواجهها صعوبات في ترجمة العناصر السياقية والثقافية للنص، وتعتمد جودتها على نوع النظام وكيفية تدريبه، إلا أنها فعالة للغاية بالنسبة لأنواع معينة من المحتوى وحالات الاستخدام، التي تشمل على سبيل المثال، النصوص المتكررة، واللغة المنظمة، وغيرها الكثير.

قبل عام 2014 كانت الخوارزميات الأساسية المستخدمة في مشاريع الترجمة الآلية من اللغة الانكليزية إلى اللغة الفرنسية هي :

**2-1-خوارزميات الترجمة ما قبل 2014 :**

**2-1-1- Knowledge-based algorithms**: تستخدم هذه الخوارزمية مجموعة من القواعد اللغوية والقواعد النحوية لتحليل النص الأصلي وتوليد الترجمة المناسبة. ومن الأمثلة على هذه الخوارزميات الشهيرة هي SYSTRAN وLOGOS .

**2-1-2- Statistical Machine Translation :** تستخدم الترجمة الآلية الإحصائية نماذج الترجمة الإحصائية الناتجة عن تحليل بيانات التدريب أحادية وثنائية اللغة. ويستخدم هذا النهج أساسا القدرة الحاسوبية لبناء نماذج بيانات معقدة لترجمة لغة مصدر إلى لغة أخرى. يتم اختيار الترجمة من بيانات التدريب باستخدام خوارزميات لتحديد الكلمات أو العبارات الأكثر تواترا , يعتبر بناء نماذج الترجمة الآلية الاحصائية عملية سريعة وبسيطة نسبيا , تستخدم هذه التقنية وحدة المعالجة المركزية بشكل مكثف، وتتطلب تكوين أجهزة واسعة لتشغيل نماذج الترجمة بمستويات أداء مقبولة.

**2-2-خوارزميات الترجمة ما بعد 2014 :**

مننذ عام 2014 تم تطوير العديد من الخوارزميات الجديدة في مجال الترجمة الآلية من اللغة الانكليزية إلى اللغة الفرنسية , أصبحت تعتمد على Neural Network , Deep learning , من بين هذه الخوارزميات:

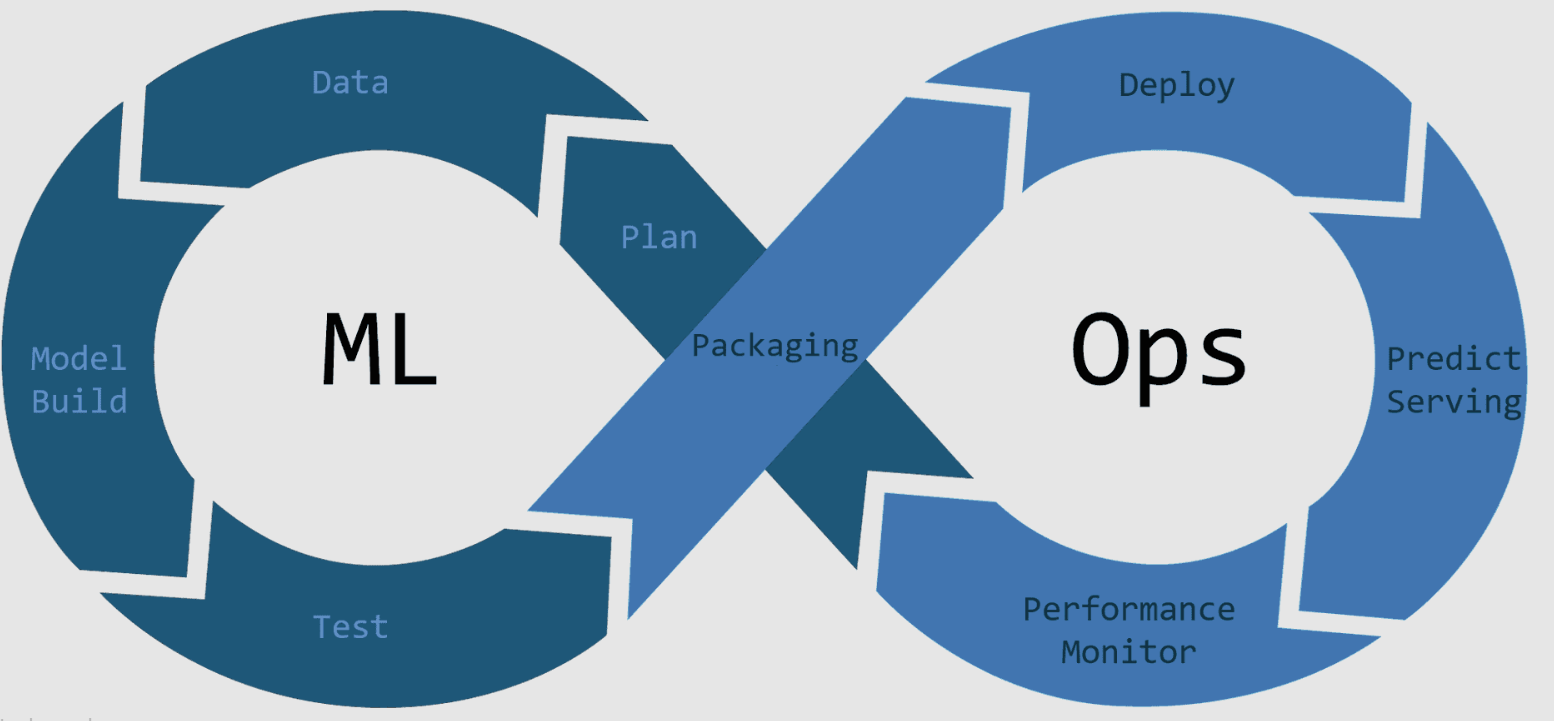
**2-2-1**- **Neural-based :** تستخدم هذه الخوارزمية شبكات عصبية لتحليل النص الأصلي وتوليد الترجمة المناسبة. وقد بدأ استخدام هذه الخوارزمية في مشاريع الترجمة الآلية منذ العام 2014، ومن الأمثلة الشهيرة على هذه الخوارزميات هي Google Neural Machine Translation وMicrosoft Translator .

**2-2-2- Deep Learning-based Neural Machine Translation :** وهي خوارزمية تستخدم شبكات عصبية عميقة لتحليل النص الأصلي وتوليد الترجمة المناسبة. ومن الأمثلة الشهيرة على هذه الخوارزميات هي BERT وGPT-3 .

**2-2-3- Artificial Intelligence-based Models :** وهي خوارزمية تستخدم تقنيات الذكاء الاصطناعي الأخرى بالإضافة إلى التعلم العميق لتحليل النص الأصلي وتوليد الترجمة المناسبة. ومن الأمثلة الشهيرة على هذه الخوارزميات هي IBM Watson Language Translator .

**3-الفصل الثالث: الجزء العملي**

تحدثنا في فقرة منهجية البحث عن عملية الترجمة و التي تبدأ بعملية المعالجة وتنتهي بمرحلة الحصول على model قادر على تحويل الجملة الانكليزية إلى فرنسية ولكن في هذا الفصل سوف نتحدث عن مفهوم أشمل وهو MLOps cycle و هي باختصار دورة حياة نموذج التعلم الآلي من مرحلة التخطيط plan إلى مرحلة الأنتاج deployment والذي يوضحه الشكل التالي



**المرحلة الأولى: التخطيط planning**

في هذه المرحلة تم العمل على :

* تقسيم أعضاء الفريق إلى عدة فرق:

1. فريق البحث (Research Team) المسؤول عن وضع معمارية النماذج و التحكم بمعاملات التدريب وتدريب النماذج ومراقبة تدريب النماذج.
2. فريق (MLOps Team) المسؤول عن التعامل مع جميع الفرق وتأمين متطلباتهم من توفير أدوات لجعل مرحلة التدريب automation و مراقبة النماذج monitor والعمل على إيصال النماذج إلى مرحلة production.
3. فريق الاختبار (Testing Team) المسؤول عن اختبار النماذج والتأكد من كفاءتها.

* تقسيم بيئة العمل إلى بيئتين:

1. بيئة التطوير والاختبار (Dev/Test)، وهنا سيتم تقسيم هذه المرحلة إلى ثلاث مراحل

بيئة التدريب train وهي تتم على سيرفر التدريب الذي يحوي GPU وعلى اعتبار لا يوجد سيرفر مستقل لدينا, تم اعتبار سيرفر colab هو السيرفر الخاص بعملية التدريب.

بيئة التطوير dev وهي تتم على سيرفر مستقل (أجهزة الحاسب الخاص بنا) حيث يتم تحويل الأوزان الناتجة عن مرحلة التدريب إلى (Web Service).

بيئة test وهي تتم على نفس سيرفر colab حيث يتم فيها اختبار النماذج الجديدة واختبار عمل function الخاصة بمعالجة الداتا.

1. مرحلة الإنتاج (Prod) حيث يتم فيها نشر (Web Service) في بيئة الإنتاج واختبارها وتحسينها

* التركيز على معمارية المشروع وفقا لأفضل الممارسات في السوق الحالية, وتم
* تتطلب معمارية المشروع الاهتمام بعدة ملفات، منها ملف خاص بتدريب النموذج (Model Training)، وملف خاص بطرق لمعالجة البيانات (Data Processing)، وملف خاص بخطوات الاختبار والتحقق من صحة المعالجة والنموذج (Testing and Verification)، وأيضاً ملف خاص بمجلد FastAPI الخاص بتحويل النموذج إلى خدمة خلفية (Backend Service) باستخدام FastAPI. يتطلب كل من هذه الملفات التفكير في التفاصيل الدقيقة والتأكد من أن كل وظيفة تعمل بشكل مستقل بالإضافة إلى التأكد من توافقها مع بعضها البعض.

**3-2 النتائج :**

**3-3 الصعوبات :**

**3-4 التوصيات :**

**المراجع Refrences :**