



مشروع

تحليل تغريدات اللغة العربية

من خلال تحليلها وتصنيفها

فريق العمل

رهف العيسى / داليا العنزي / أمل الشهابي

مبادرة سدايا

من ضمن مبادرات سدايا التي تسعى من خلالها إلى التركيز على معالجة البيانات وتقديم خدمات وحلول تساهم في تطبيق أعلى استفادة من البيانات و أدوات الذكاء الاصطناعي، ومن ضمن هذه المبادرات مبادرة التركيز على تقديم حلول في معالجة البيانات المتعلقة باللغة العربية، ومن هذا المنطلق سعينا من خلال دورنا كعلماء بيانات على تسليط الضوء على اللغة العربية.

من الألف إلى الياء

نطور، ونستحدث، ونبتكر

• طموحنا •

أن تكون المملكة العربية السعودية الرائدة عالمياً في مجال معالجة اللغة العربية بالذكاء الاصطناعي

وضعت سدايا خدمة اللغة العربية كأحد أولوياتها وذلك عبر:

تأسيس مركز التميز لمعالجة اللغة العربية بالذكاء الاصطناعي بالشراكة مع عدد من الجهات الوطنية ذات العلاقة والمؤسسات الأكاديمية والشركات العاملة الرائدة في هذا المجال.	تطوير إستراتيجية معالجة اللغة العربية بتقنيات الذكاء الاصطناعي، التي تضمنت أهدافاً ومبادرات إستراتيجية لتتبعها المملكة الريادة في المجال.
تطوير نماذج التعرف الصوتي على الرموز والنصوص العربية باستخدام تقنية الرؤية الحاسوبية.	تطوير نموذج متقدم للتعرف التلقائي على الكلام باللغة العربية، مع التركيز على اللهجات السعودية، وذلك بنسبة دقة تصل إلى 96%.

عن معالجة اللغة الطبيعية:

تعد معالجة اللغة الطبيعية مجالاً فرعياً من الذكاء الاصطناعي، وهو العلم الذي يعني جعل الآلة قادرة على فهم وتوليد اللغة مثل البشر. ومجرد أن نفهم اللغة ما نوله ونكتبه، يمكننا بعد ذلك بناء تطبيقات قيمة عن طريق استخراج المعنى والمعلومات من المدخلات الصوتية أو النصية.

SDAIA
الهيئة العامة للغذاء والدواء
Saudi Drug Authority

SDAIA.SA @SDAIA_SA

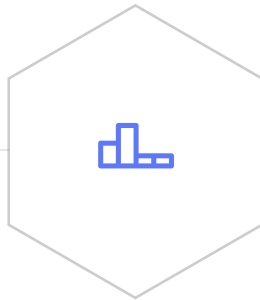
صورة من منشور المبادرة *

فكرة المشروع

في طور عملنا على هذا المشروع بحثنا عن أكثر المصادر التي تحتوي على النص العربي في مواقع التواصل الاجتماعي و كان أبرز تلك المنصات هي منصة **تويتر** التي تحتوي على ملايين المنشورات باللغة العربية في عدة مجالات متنوعة، ومن هذا المنطلق قررنا العمل على الاستفادة من تلك البيانات من خلال سحبها و تحليلها ومن ثم تصنيفها باستخدام أدوات الذكاء الاصطناعي.

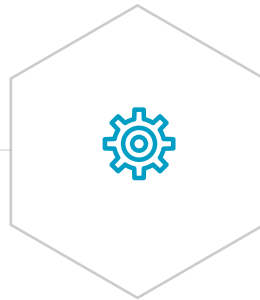


مالذي قمنا بعمله؟



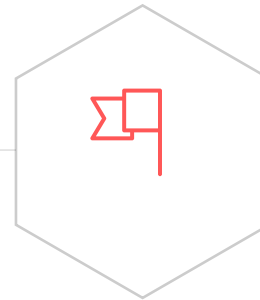
مرحلة استعراض البيانات

في هذه المرحلة قمنا باستعراض البيانات عن طريق الرسم البياني



مرحلة معالجة البيانات

قمنا في هذه المرحلة بعدة مراحل لمعالجة البيانات وتحليلها



مرحلة سحب البيانات

في هذه المرحلة قمنا بسحب API البيانات من خلال تويتر



البداية



مالذي قمنا بعمله؟



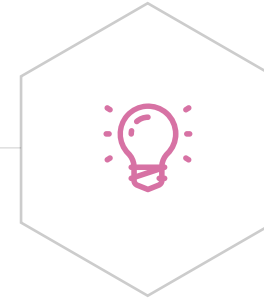
مرحلة بناء لوحة البيانات

في هذه المرحلة قمنا بعمل داشبورد توضيحي يعكس نتائج التي توصلنا لها في تحليل البيانات



مرحلة التحقق من النتائج

قمنا في هذه المرحلة بعملية اختبار للتحقق من نتائج المودل المختار



Model مرحلة بناء الـ

في هذه المرحلة قمنا بتجربة أكثر من نوع للمودل تعلم الآلة



مرحلة سحب البيانات

٢ استخدام وتوليد API's keys
البدأ بالكودينق

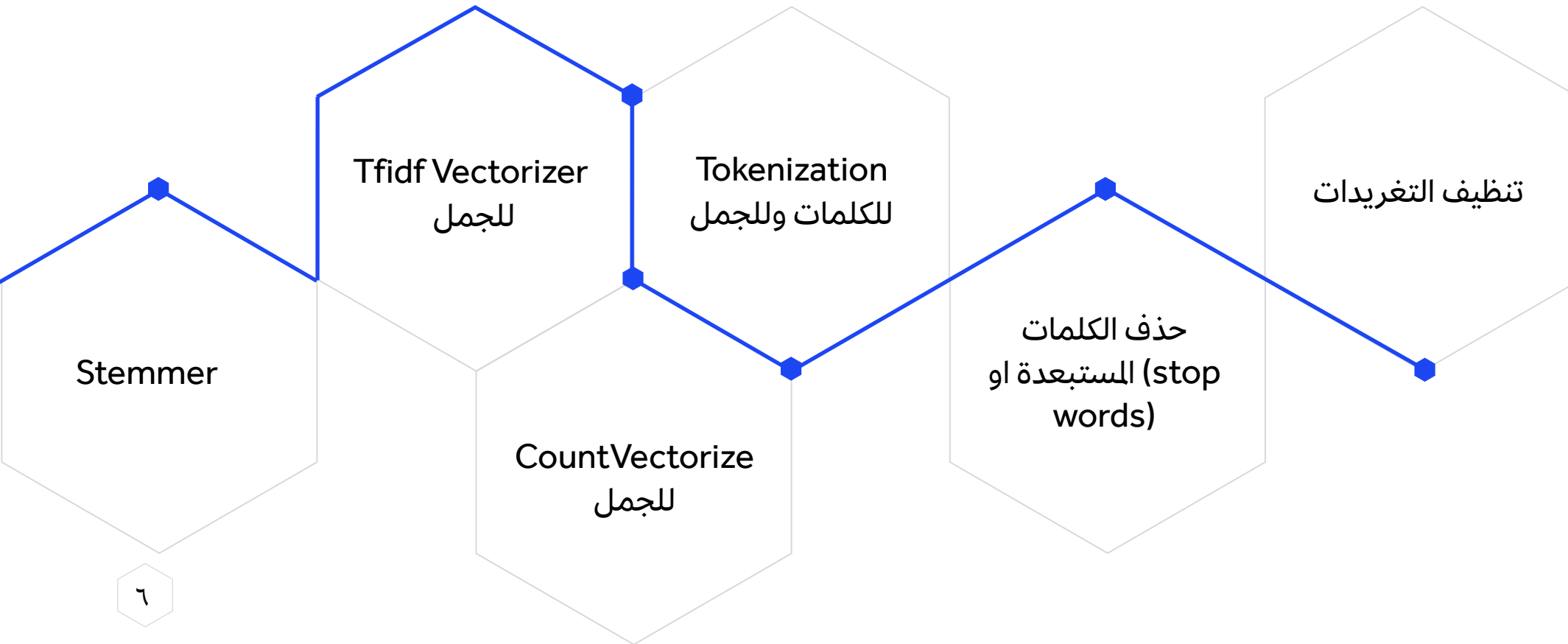
١ سحب الداتا قانونيًا
رفع طلب حساب المطورين (developer account)

٤ حفظ الداتا في ملف اكسل

٣ تصفية محتوى التغريدات



مرحلة معالجة البيانات





مرحلة استعراض البيانات

EDA



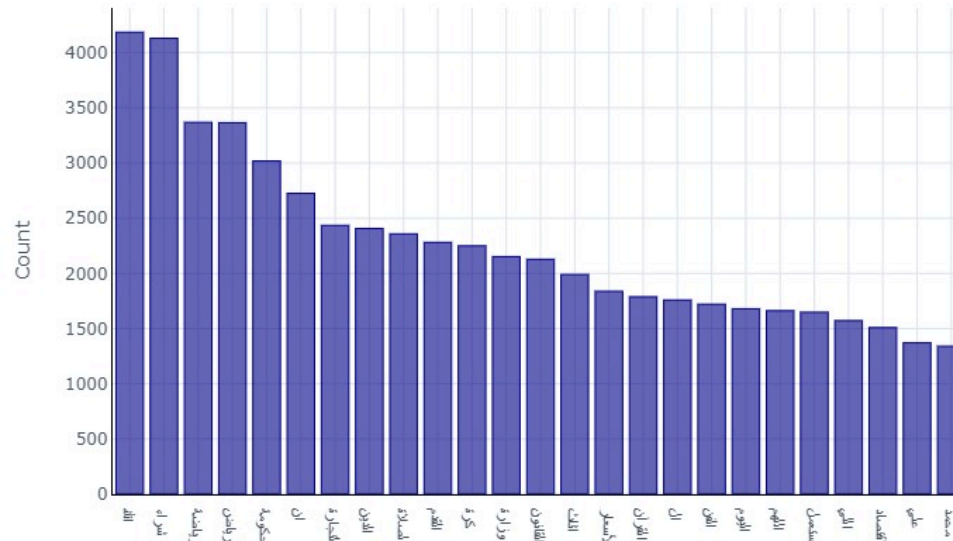
١ - رسم بياني يستعرض في تصنيفات التغريدات.



مرحلة استعراض البيانات

EDA

أعلى 20 كلمة تكرارا في التغريدات



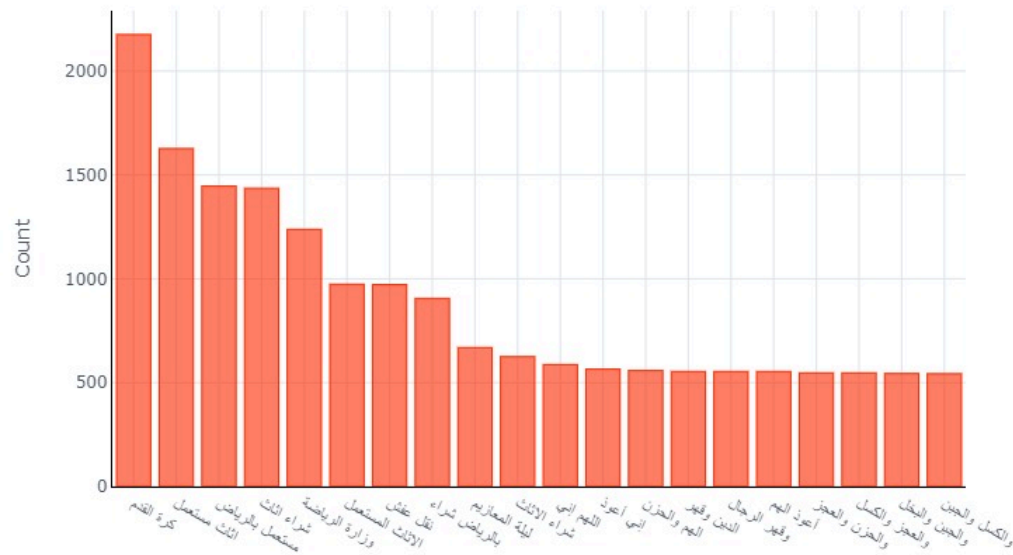
٢- رسم بياني يستعرض فيه أكثر الكلمات تكرارا في التغريدات.



مرحلة استعراض البيانات

EDA

أعلى 20 كلمات تالية تكرارا في التغريدات



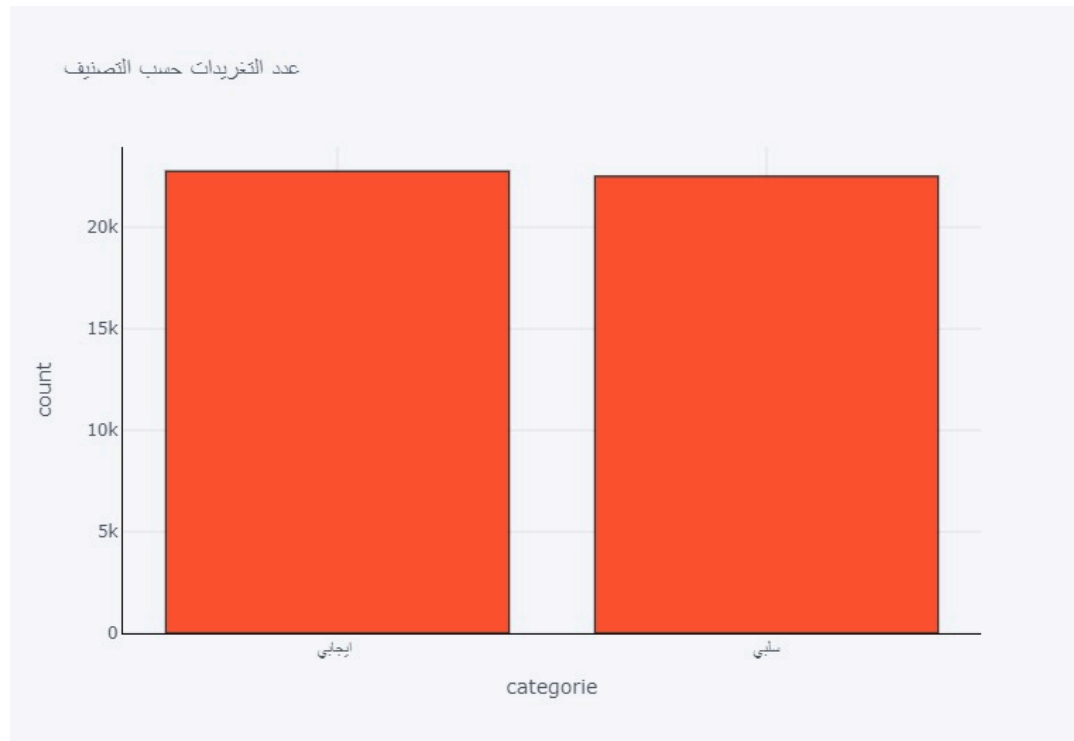
٣- رسم بياني يستعرض فيه أكثر الكلمات تكرارا في التغريدات.





مرحلة استعراض البيانات

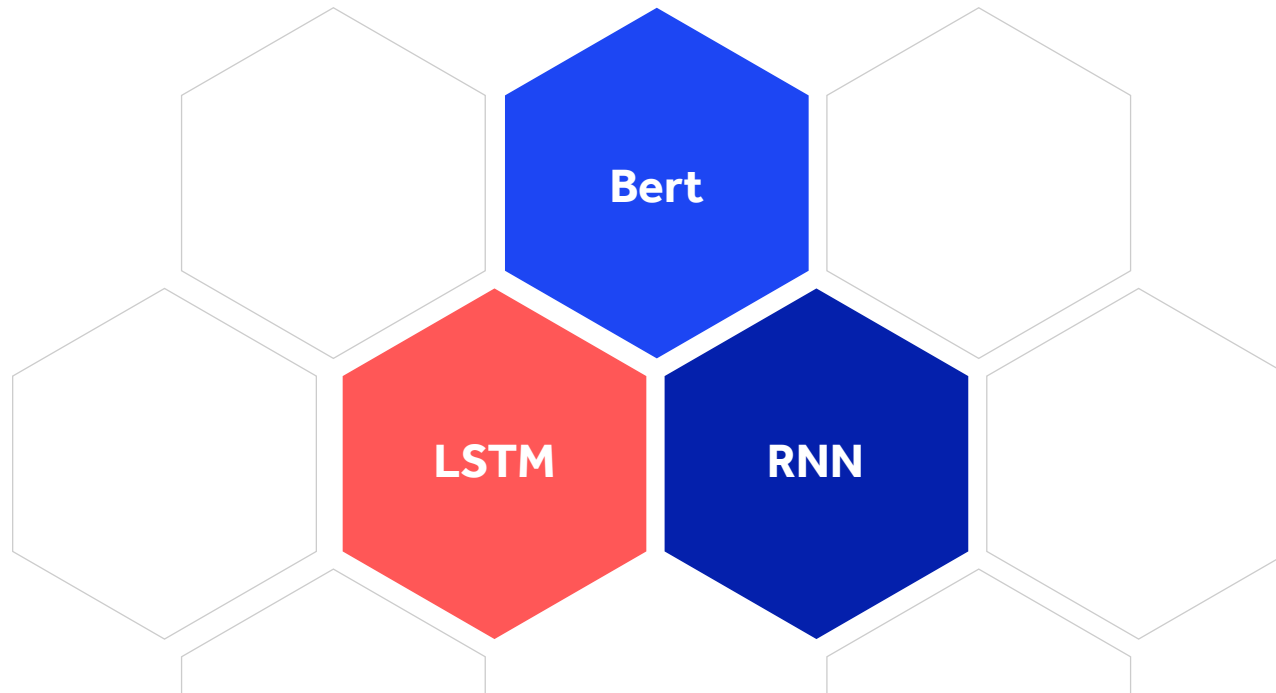
EDA



٤- رسم بياني يستعرض فيه تحليل المشاعر .



مرحلة بناء Model





Bert مودل

Report Bert Model لتحليل المشاعر

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.77	0.82	0.80	4514
---	------	------	------	------

1	0.81	0.76	0.79	4535
---	------	------	------	------

accuracy			0.79	9049
----------	--	--	------	------

macro avg	0.79	0.79	0.79	9049
-----------	------	------	------	------

weighted avg	0.79	0.79	0.79	9049
--------------	------	------	------	------

```
array([[3724, 790],  
       [1083, 3452]])
```



Bert مودل

Report Bert Model للتصنيفات

precision recall f1-score support

0	0.94	0.95	0.95	1891
1	0.96	0.95	0.95	1803
2	0.93	0.93	0.93	1894
3	0.95	0.95	0.95	1855
4	0.96	0.97	0.96	1974

accuracy		0.95	9417	
macro avg	0.95	0.95	0.95	9417
weighted avg	0.95	0.95	0.95	9417

```
array([[1796, 11, 27, 33, 24],  
       [ 23, 1708, 42, 21, 9],  
       [ 26, 43, 1754, 37, 34],  
       [ 28, 10, 36, 1768, 13],  
       [ 29, 7, 17, 8, 1913]])
```



مودل RNN

RNN Model للتصنيفات

Accuracy: 20%

مودل LSTM

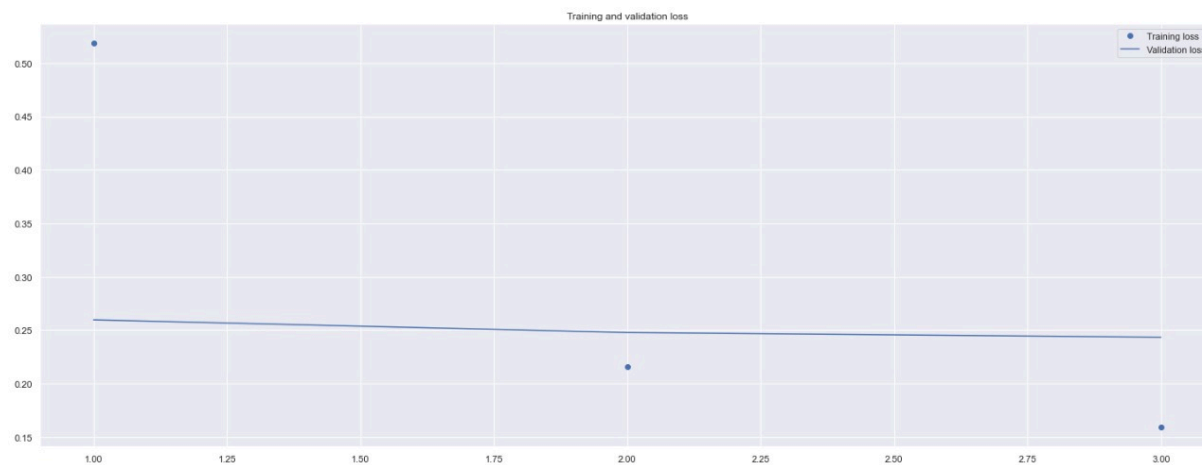
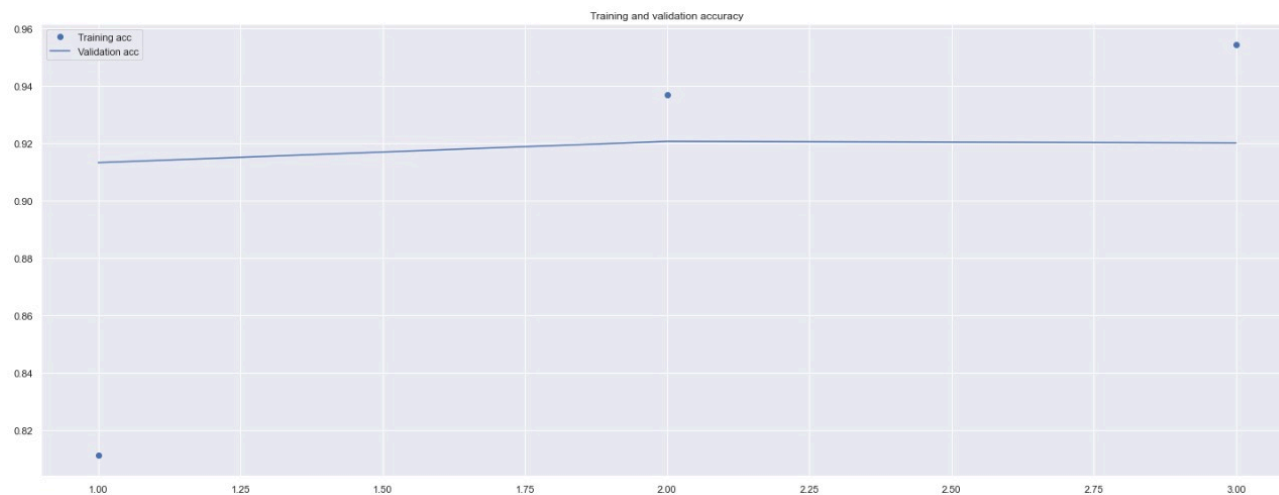
LSTM Model للتصنيفات

Accuracy: 91%



مودل LSTM

رسم بياني لمودل LSTM





```
array([[5724, 790],  
       [1083, 3452]])
```

Let's make a prediction

```
✓ [26] p = ktrain.get_predictor(learner.model, t)
```

```
▶ p.predict(["|"])
```

```
[50] p.predict("")
```

▼ Saving the model

مرحلة التحقق من النتائج

Demo

فديو توضيحي على تجربة المودل والتحقق منه
في تحليل المشاعر



```
[ 30, 4, 24, 7, 1909]])
```

Let's make a prediction

```
✓ [20] p = ktrain.get_predictor(learner.model, t)
```

```
▶ p.predict(["|"])
```

```
[ ]
```

▼ Saving the model

مرحلة التحقق من النتائج

Demo

فديو توضيحي على تجربة المودل والتحقق منه
في التصنيف



تعامل ديب ليرنيق
مودل مع اللغة
والعربية والمشاكل

التعامل مع
اللغة العربية
وقلة المراجع

تحدي سحب الداتا

التحديات التي واجهتنا

في المشروع وقدردنا على التغلب عليها



مرحلة بناء لوحة البيانات





الختامية

الحمد لله على تمام المشروع بنجاح

في العمل المشتقيل
نتمنى تطوير وأنشاء
موقع خاص في تحليل
وتصنيف آراء ومشاعر
الناس بالنص العربي

شكرا لكم

نتمنا ان يكون نال اعجابكم مشروعنا
وما تم طرحه

فريق العمل
رهدف العيسى / داليا العنزي / أمل الشهابي

