



الجمهورية العربية السورية
جامعة الشام الخاصة
كلية الهندسة المعلوماتية

تقرير عن Features extraction using fine-tuned BERT
العام الدراسي 2023-2024

Features Extraction Using Fine-tuned BERT

المحتوى

3	مقدمة
3	Fine-tuned
3	عملية ال Fine-tuning
3	النماذج المستخدمة للتصنيف
3	الطريقة
3	النتائج
4	استنتاجات
4	تطلعات

مقدمة

كما هو معلوم في مسائل التصنيف عادة يتم استخدام النماذج مسبقاً للتدريب بطريقتين إما إجراء عملية Fine-tuned للنموذج واستعماله مثل ما تم تطبيقه عند تطوير ال Baseline أو استخدامه للقيام ب Features Extraction واستخدام رأس تصنيف (نماذج تعلم آلي أو شبكات عصبونية) للقيام بعملية التصنيف، ففي هذا التقرير نقدم التوجه الثاني ونعرض النتائج التي توصلنا إليها.

Fine-tuned

لقد اعتمدنا في تطبيق ال Features Extraction على استخدام نموذج AraBERT وطبقنا ذلك بطريقتين:

- 1- استخدام نموذج AraBERT خام (من دون إجراء عملية Fine-tuned له).
- 2- استخدام نموذج AraBERT قمنا بتدريبه مسبقاً.

عملية ال Fine-tuning

قمنا بتدريب النموذج من خلال إضافة خطية تقوم بأسقاط خرج النموذج الى خرجين أحدهما يمثل ان المراجعة حقيقة والأخرى يمثل ان المراجعة مخادعة، فقمنا بتقسيم البيانات الى بيانات تدريب وبيانات اختبار بنسبة 20/80 وهي نفس التقسيمة التي استخدمناها في جميع النماذج لضمان ان بيانات التدريب هي بيانات لم يسبق للنموذج ان رآها من قبل وبعد إتمام التدريب نقوم بإزالة الطبقة الأخيرة ونستعمل النموذج للقيام بعملية ال Features Extraction من اجل نماذج التصنيف التي سنستخدمها.

النماذج المستخدمة للتصنيف

تم استخدام 12 نموذج للقيام بعملية التصنيف مع نموذج ال AraBERT ك Features Extractor، النماذج هي:

- 1- 2-Layer NN
- 2- KNN Classifier
- 3- Decision Tree Classifier
- 4- Extra Tree Classifier
- 5- Random Forest Classifier
- 6- Linear SVC
- 7- SVM Classifier
- 8- NuSVC
- 9- SGD Classifier
- 10- Logistic Regression Classifier
- 11- Logistic Regression Classifier with CV
- 12- Gaussian Naïve Bayes Classifier

الطريقة

نقوم بتحميل نموذج AraBERT سواء الخام أو الذي تم تدريبه لكي نحول كل من بيانات التدريب والاختبار من نصوص الى تمثيل رقمي (embedding)، من اجل النماذج نستخدم مكتبة Sklearn ونقوم بتدريب النماذج على البيانات المحولة ومن ثم نقوم باختبار هذه النماذج على بيانات الاختبار المحولة.

النتائج

نشرح في الجدول ادناه الدقة ومعيار f1 لجميع النماذج بالطريقتين المطروحتين أعلاه.

Table 1

Model Name	Fine-tuned BERT		Raw BERT	
	Accuracy	F1	Accuracy	F1
2-Layer NN				
KNN Classifier	0.93	0.93	0.72	0.74
Decision Tree Classifier	0.92	0.92	0.64	0.65

Extra Tree Classifier	0.92	0.92	0.60	0.60
Random Forest Classifier	0.94	0.94	0.75	0.75
Linear SVC	0.93	0.94	0.76	0.77
SVM Classifier	0.94	0.94	0.78	0.79
NuSVC	0.93	0.93	0.80	0.81
SGD Classifier	0.94	0.94	0.73	0.73
Logistic Regression Classifier	0.94	0.94	0.78	0.79
Logistic Regression Classifier with CV	0.94	0.94	0.77	0.78
Gaussian Naïve Bayes Classifier	0.92	0.92	0.66	0.71

استنتاجات

من الجدول أعلاه نلاحظ كما هو متوقع ان نتائج النماذج التي تستخدم النموذج المدرب مسبقا تكون أفضل، وكذلك بالمقارنة مع توجه ال Fine-tuned نلاحظ ان توجه ال Features Extraction مع نماذج تعلم الي للتصنيف تعطي نتائج مستقرة حيث مجال الدقة [0.91-0.94] وهو مجال متقارب اما بالنسبة لتوجه ال Fine-tuning فالمجال كان كبير [80-91] ويمكن القول ان أحد أسباب ذلك هو ان حجم البيانات ليس كبير بما فيه الكفاية لتدريب نموذج تعلم عميق وهذا ليس الحال بالنسبة لنماذج التعلم الآلي التي تطلب كمية بيانات اقل.

تطلعات

على الرغم من ان الدقة ومعيار ال F1 عاليا وليس هنالك تذبذب بمجالاتها ومع دراسة البيانات المستخدمة الا انه ما زالت ال Features المستخدمة للتصنيف ليست واضحة لذلك كان من الضروري انشاء مجموعة بيانات تكون معايير التصنيف فيها واضحة لكي نقدر على تدريب واختبار النماذج لدينا بشكل واقعي وأفضل.