**Lecture 3**

* **Mention Two Ways for Feature Engineering?**

**Classical NLP Models**

**A picture containing text, screenshot, diagram, rectangle

Description automatically generated**

**, DL-NLP Models**

**A diagram of a deep learning

Description automatically generated with low confidence**

* **Explain How to Do Modeling?**

**Human heuristics can provide a good start.**

**Ex. We may have a blacklist for domain that are used to send spam.**

**قائمة سوداء بأسماء النطاقات المواقع التي ترسل الرسائل الالكترونية المزعجة**

* **How we can extract information such as email iDs, dates, telephone with …. Re (regular expressions)**
* **…….,…… used for defining advanced regular expressions. Stanford NLP’s, TokensRegex and spaCy’s**
* **What is the disadvantage of adding more heuristics?**

**Complex rule-based Model, harder to maintain, understand best way to combine heuristic with ML :**

* + **Create a feature from the heuristic.**
  + **Pre-process your input to the ML Model**

**يعنى الموديل بيكون معقد لان مبنى على قواعد كتير جدا وده هيخلى في صعوبة الصيانة وصعوبة في الفهم**

**احسن حاجة انى ادمجهم الطريقتين مع بعض**

* **Mention Approaches for building the model.**

**1-Stacking**

**Model stacking**

**collection of ML models**

**model stacking involves training multiple models and then using their predication as input to another meta-model which learns how to combine the predictions.**

**Model Ensembling**

**involves combining the predictions of multiple models to improve the overall prediction.**

**A diagram of a stacking process

Description automatically generated with medium confidence**

**Better feature engineering**

1. **Feature Extraction: يتمثل في استخراج مجموعة من الميزات أو السمات من البيانات النصية. يتم استخدام هذه الميزات كمدخلات لتدريب نماذج التعلم الآلي. من بين البارامترات الهامة في هذه المرحلة هي طريقة تمثيل النص (مثل ترم-فريكوينسي) وحجم الميزات (عدد الكلمات أو العبارات المختلفة).**
2. **Feature Selection: يتمثل في اختيار مجموعة محددة من الميزات الهامة من بين الميزات المستخرجة في المرحلة الأولى. يتم استخدام هذه الميزات المختارة كمدخلات لتدريب نماذج التعلم الآلي. من بين المعايير الهامة في هذه المرحلة هي القدرة على التمييز بين المجموعات المختلفة من النصوص والحد من الأبعاد الزائدة.**
3. **Feature Reduction: يتمثل في تقليل عدد الميزات المستخدمة في تدريب نماذج التعلم الآلي. يتم استخدام هذه المرحلة في حالة وجود عدد كبير جدًا من الميزات التي يصعب معالجتها بشكل فعال. من بين التقنيات الشائعة في هذه المرحلة هي تقنيات تقليل الأبعاد مثل PCA و SVD.**

**Transfer learning:**

**Transfer preexisting knowledge from a big, well-trained model to a newer model at its initial phase (pre training)**

**Without losing general features**

* **يتضمن التعلّم النقلي نقل المعرفة الموجودة مسبقًا في موديل مدرب بطريقة جيدة والذي تم تدريبه على مجموعة بيانات كبيرة، وتطبيقها على موديل جديد في مرحلة ابتدائية من التدريب. يتم ذلك لتوفير قيم ابتدائية للأوزان في الموديل الجديد.**

**The new model slowly adapts to the task at hand(fine-tuning).**

* **يتأقلم الموديل الجديد ببطء مع المهمة الجديدة المطلوبة منه لأنه يحتاج إلى ضبط المعرفة الموجودة مسبقًا لتتناسب أكثر مع المهمة الجديدة. يتم ذلك عن طريق ضبط أوزان الموديل خلال التدريب، مع الحفاظ على الميزات العامة التي تم تعلمها من المعرفة الموجودة مسبقًا. يجب أن يتم عملية التضبيط ببطء وبحرص لتجنب التدريب الزائد أو فقدان الميزات العامة التي تم تعلمها من المعرفة الموجودة مسبقًا.**

**Traditional ML Vs Transfer Learning**

**A picture containing text, screenshot, diagram, circle

Description automatically generated**

**Reapplying Heuristic**

**ML Models still make mistakes.**

**Review at the end of the modeling pipeline to find any common pattern in errors and correct them**

**لنفترض أن لدينا نموذج تنبؤي لتحديد ما إذا كان العملاء سيشترون منتجًا معينًا أم لا، وقد تم تدريب هذا النموذج باستخدام مجموعة بيانات كبيرة من العملاء. ومع ذلك، يتبين أن النموذج يرتكب بعض الأخطاء في التنبؤ بالعملاء الذين سيشترون المنتج.**

**بعد إجراء مراجعة في نهاية سير النمذجة، تم العثور على أنماط مشتركة في الأخطاء التي يرتكبها النموذج. ومن خلال هذه المراجعة، تم اكتشاف أن النموذج يرتكب أخطاء في التنبؤ بالعملاء الذين يعيشون في مناطق معينة أو العملاء الذين يتبعون أنماط شرائية معينة.**

**بناءً على هذه المعرفة، يمكن تعديل النموذج لتصحيح هذه الأخطاء، على سبيل المثال، يمكن استخدام معلومات حول المناطق الجغرافية أو أنماط الشراء لتحسين توقعات النموذج وتصحيح الأخطاء الشائعة.**

**وبهذا الإجراء، يمكن تحسين دقة التنبؤات وأداء النموذج في المستقبل، وتحسين قدرته على تحديد العملاء الذين سيشترون المنتج.**

**Apply domain-specific knowledge that is not automatically captured in the data during the training.**

**بمعنى انه يمكن تطبيق المعرفة الخاصة بمجال معين والتي لا يتم التقاطها تلقائيا من البيانات المستخدمة اثناء التدريب**

**Choose the model based on data availability.**

**\*No Data: we fully rely on heuristics**

**\*a small amount of data: active learning, transfer,…..**

**\*a lot of data: a range of modeling techniques**

**Mean of Active learning:**

**لنفترض أن لدينا مجموعة كبيرة من الصور للأشخاص، ونريد تدريب نموذج لتحديد ما إذا كانت الصور تحتوي على شخص أم لا. ولكن، لا يوجد لدينا الوقت أو الموارد الكافية لتسمية كل الصور يدويًا.**

**بدلاً من ذلك، يمكن استخدام Active Learning لتحديد عينة صغيرة من الصور الغير مصنفة التي يتم تسميتها يدويًا بواسطة محلل بشري. بعد ذلك، يتم استخدام هذه التسميات الجديدة لتحديث النموذج، ثم تم استخدام النموذج لتصنيف أكبر عدد من الصور المتبقية، واختيار عينة جديدة من الصور الغير مصنفة لتسميتها يدويًا.**

**وبهذه الطريقة، يتم تحديد الصور الأكثر معلوماتية والأكثر فائدة لتحسين أداء النموذج، وبالتالي يمكن تحديث النموذج بشكل فعال بأقل جهد وتكلفة ممكنين، وبناء نموذج أكثر دقة في تحديد وجود الأشخاص في الصور.**

**A picture containing text, screenshot, font, number

Description automatically generated**

**we measure the performance on unseen data**

**Evaluation Should?**

**1-use the right metric for evaluation**

**2-follow the right evaluation process**

**What is the two types of Evaluation?**

**\*Intrinsic (intermediary objective)**

**The output of the NLP model on a data point is compared**

**Against the corresponding label for that data point and**

**Metrices are calculated based on match or mismatch.**

**Between output and label (****binary ,one-to-words(one to many), large text itself)**

**A picture containing text, screenshot, font, number

Description automatically generated**

**\*Extrinsic (final objective)**

**Evaluate the model performance on the final objective.**

**Regression model is build to rank emails and bringing the mot important emails to the top of the inbox consider the model does well on the ML metrics but doesn’t really save a lot of time users**

**يتم قياس الأداء بناء على الهدف النهائي من الموديل فعلى سبيل المثال موديل يستخدم لتصنيف كل رسالة تبعا لأهميتها ولكن هذا الموديل يستغرق وقت طويلا ولا يوفر وقت فهو لم يحقق الهدف النهائي للأسف**

**A picture containing text, screenshot, font, number

Description automatically generated**

**A picture containing text, screenshot, font, number

Description automatically generated**