**Lecture 4**

* **Automatic Vs Manual Evaluation?**

**Automatic:**

* + **Comparing the system’s output with the gold standard output**

**معناها انى بقارن الاوت بوت الخارج من السيستم مع الاوت بت اللى المفروض يطلع واللي اتعمل يدويا من خلال المستخدم**

* + **The cost of producing the gold standard time-consuming**

**الموضوع اول مرة بياخد وقت كبير لان بعمل الموضوع يدويا وعلى داتا سيت كبيرة**

* + **But easily repeatable without additional cost**

**بمجرد انى عملت المعيار الذهبي (الاوت بوت ) ممكن بعدها استخدمه اكتر من مرة عشان اعمل اقيس الأداء الخاص بالسيستم بتاعى**

**Manual:**

* + **For many nlp problems the definition of a gold standard can prove impossible (when inter-annotator agreement insufficient)**

**للعديد من المشكلات من الصعب الحصول على الاوت بوت الصحيح (الذهبى) عند يكون اتفاق المعلقين مش كفاية**

* + **Manual evaluation is performed by human judges**

**اليدوي بيتم عن طريق احكام الانسان**

* **Intrinsic Vs Extrinsic Evaluation**

**Intrinsic:**

**(intermediary objective) consider an isolated NlP system and measure its performance mainly with respect to a gold standard result (precision, recall ,f1 ,accuracy)**

**بيعتبر ان الجزء اللى معمول بمعالجة اللغة الطبيعية جزء معزول ويتم قياس الأداء الخاص بى عن طريق الجولد استاندرد واللى هو الاوت بوت**

**يتم تقييم أداء النظام بناء على مقاييس داخلية وليس بناء عل**

**Extrinsic(final Objective) consider the nlp system as a component in a more complex setting the amount of time users spent on a spam email بيعتبر السيستم الخاص بالمعالجة الطبيعة جزء من اعدادات اكبر معقدة على سبيل المثال الوقت اللى بيخده المستخدم على ايميل مزيف**

**For example, imagine a situation where an NLP system is designed to detect phishing emails, but it fails to identify a significant number of them. In this case, users may spend a lot of time trying to determine whether an email is legitimate or not, which can put them at risk of falling victim to a phishing attack** **على سبيل المثال، تخيل حالة يتم فيها تصميم نظام معالجة لغوية طبيعية (NLP) للكشف عن رسائل الصيد الاحتيالي (Phishing)، ولكن يفشل في التعرف على عدد كبير من تلك الرسائل. في هذه الحالة، قد ينفق المستخدمون الكثير من الوقت في محاولة تحديد ما إذا كانت رسالة مشروعة أم لا، مما يعرضهم لخطر الوقوع ضحية هجوم صيد الأسماك الاحتيالي.**

**الخلاصة**

**في حالة التقييم الداخلي، يتم تقييم أداء النظام بناءً على مقاييس داخلية مثل دقة التصحيح الإملائي وليس بناءً على تأثيره على المستخدمين في العالم الحقيقي، وهو ما يميزه عن التقييم الداخلى (Extrinsic Evaluation).**