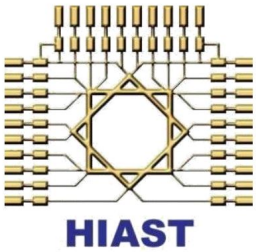
الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم المعلوميات
العام الدراسي 2017/2018



تقييم جودة نصوص اللغة الانكليزية

مشروع السنة الرابعة

إعداد
فاروق حجابو

إشراف
د. غيداء ربداوي م. رياض سنبل

2 آب 2018

الملخص

المحتويات

قائمة الأشكال

قائمة الجداول

الاختصارات

SVM	Support Vector Machine
SMO	Sequential Minimal Optimization
NLP	Natural Language Processing

المصطلحات

Artificial Intelligence	الذكاء الصنعي
Machine Learning	تعلّم الآلة
Natural Language Processing	معالجة اللغات الطبيعية
Supervised Learning	التعلم تحت الإشراف
Unsupervised Learning	التعلم بدون إشراف
Semi-Supervised Learning	التعلم نصف المشرف عليه
Reinforcement Learning	التعلم بالتعزيز
Classification	التصنيف
Regression	الانحدار
Training Set	معطيات التدريب
Test Set	معطيات الاختبار
Training Instance	مثال تدريبي
Accuracy	الصِحَّة
Precision	الدقّة
Recall	الإرجاع
Clustering	التجميع
Features	ميزات
Feature Extraction	استخراج الميزات
Regularization	التنظيم
Kernel	نواة

Linear Kernel	النواة الخطية
Polynomial Kernel	النواة الحدودية
Gaussian Kernel	النواة الغاوسية
Hyperparameter	باراميتر فوقي
Classifier	مُصنّف
Morphological Analysis	التحليل الصرفي
Tokenization	التقطيع
Stemming	التشذيب
Chunking	التكتيل
Parsing Tree	الشجرة النحوية
Dependency Parsing	تحليل التّبعية
Dependency Graph	بيان التبعية
Anaphora	الإحالة

الفصل الأول

التعريف بالمشروع

يُمهّد هذا الفصل للمشروع، حيث يُبيّن فكرة المشروع وأهميتها والأهداف المرجوّة منه. ويذكر المتطلبات الوظيفية وغير الوظيفية للمشروع.

1.1 مقدمة

الفصل الثاني

الدراسة المرجعية

يبيّن هذا الفصل الدراسة المرجعية للمشروع. يبدأ بتقديم مفاهيم تعلّم الآلة ومراحلها المختلفة والمعايير المعتمدة لتقييمها. ويقدّم مفاهيم ومراحل معالجة اللغات الطبيعية. وأخيراً يسرد بعض الأوراق الأبحاث العلمية المتعلقة بهذا المشروع، ويوضح المنهجيات المتبعة فيها.

1.2 تعلم الآلة

تعلم الآلة Machine Learning هو فرع جزئي من الذكاء الصنعي Artificial Intelligence. يُقصد بتعلم الآلة مجموعة الأدوات والمفاهيم والمنهجيات المستخدمة لبرمجة الحواسيب بطريقة تسمح لهذه الحواسيب بالتعلم من المعطيات [25].

ويمكن أيضاً تعريفه بشكل أكثر عمومية كالتالي:

"Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed."
–Arthur Samuel, 1959

كما يعتبر التعريف التالي تقني وأكثر دقة:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."
–Tom Mitchell, 1997

على سبيل المثال، النظام الذي يقوم بفلترة الإيميلات إلى إيميلات مؤذية spam وإيميلات غير مؤذية non-spam، يستخدم منهجيات تعلم الآلة. يقوم هذا النظام بتعلم طريقة التمييز بين هذين النوعين من الإيميلات باستخدام عدد كبير من الأمثلة والمعطيات المصنفة مسبقاً. نسمي هذه المجموعة من الأمثلة بمعطيات التدريب Training Set، وكل مثال منها نسميه مثال تدريبي Training Instance.

في هذه الحالة، المهمة T هي تصنيف الإيميلات الجديدة إلى إيميلات مؤذية وإيميلات غير مؤذية، الخبرة E هي مجموعة معطيات التدريب، ومؤشر قياس الأداء P يمكن تعريفه بعدّة طرق؛ فمثلاً يمكننا استخدام نسبة عدد الإيميلات التي تم تصنيفها بشكل صحيح إلى عدد الإيميلات الكلي (هذا المعيار يسمى الصِحَّة Accuracy كم سنرى لاحقاً).

1.1.2 تصنيفات تعلم الآلة

يمكن تصنيف أنظمة تعلم الآلة وفق عدّة معايير. التصنيف الأكثر شهرة يعتمد على آلية التدريب، وهو كالتالي:

- التعلم تحت الإشراف Supervised Learning: وهي حالة أن تكون الأمثلة التدريبية متوفرة مع الخرج label المرتبط بها. وهذه حالة مثال تصنيف الإيميلات المطروح سابقاً. حيث أن معطيات التدريب هي مجموعة كبيرة من الإيميلات المصنفة مسبقاً من قبل البشر إلى إيميلات مؤذية وإيميلات غير مؤذية.
- التعلم بدون إشراف Unsupervised Learning: وهي حالة أن تكون معطيات التدريب موجودة ولكنها غير مصنفة unlabeled أو غير مرتبطة بخرج معيّن. على سبيل المثال، قد ترغب شركة في تصنيف زبائنها إلى عدّة مستويات، زبائن من الدرجة الأولى، زبائن من الدرجة الثانية، وهكذا. فيمكن استخدام تعلم الآلة لاكتشاف بعض الأنماط الموجودة في معطيات الزبائن واكتشاف هكذا تصنيف. وهذا ما يُعرف بالتجميع Clustering.
- التعلم نصف المشرف عليه Semi-Supervised Learning: وهي حالة وسيطة بين التصنيفين السابقين. تكون فيها بعض أمثلة التدريب مرتبطة بخرج معيّن (غالباً تشكل النسبة الصغيرة)، وتكون باقي الأمثلة غير

مرتبطة بخرج. تنطبق هذه الحالة على مثال تصنيف الإيميلات في حال لم تكن جميع معطيات التدريب مصنفة بشكل مسبق.

- التعلم بالتعزيز Reinforcement Learning: وهي الحالة التي يتخاطب فيها النظام مع بيئة أخرى. تقدم له هذه البيئة نتائج feedback بناءً على أفعاله. هذا الصنف ينطبق على الخوارزميات المستخدمة لتدريب الأنظمة التي تتعلم الألعاب. حيث يقوم النظام بمجموعة من الأفعال actions ضمن بيئة اللعبة، وبناءً على النتائج (تحسّن نتيجته أو انخفاضها) يغيّر أفعاله اللاحقة.

وعلى وجه الخصوص يمكن تصنيف التعلم تحت الإشراف بحسب نوع الخرج المرتبط بمعطيات التدريب. تصنّف بشكل أساسي عريض كالتالي:

- التصنيف Classification: يكون الخرج المرتبط بكل مثال تدريبي هو صف class محدد من مجموعة صفوف. عدد هذه الصفوف قد يكون 2، 3، إلخ. في مثال تصنيف الإيميلات السابق، عدد الصفوف هو 2، حيث أن كل مثال تدريبي (إيميل معيّن من معطيات التدريب) هو إمّا مؤذي أو غير مؤذي.
- الانحدار Regression: يكون الخرج المرتبط بكل مثال تدريبي هو عدد حقيقي. مثل مسألة التنبؤ بسعر منزل بمعرفة معلومات عنه مثل مساحته، عدد الغرف، إلخ.

2.1.2 المراحل اللازمة لتطبيق تعلم الآلة

إذا عدنا إلى مثال تصنيف الإيميلات، حيث قلنا أن معطيات التدريب هي مجموعة من الإيميلات المصنفة بشكل مسبق إلى إيميلات مؤذية وإيميلات غير مؤذية. يمكن أن نسأل هنا: ما هو تحديداً الدخل؟ أي كيف سنعبّر عن الإيميل؟ بالطبع يمكن اعتبار الإيميل كنص؛ فهو مجموعة من الكلمات والرموز. ولكن كما سنرى لاحقاً، من الصعب على معظم خوارزميات تعلم الآلة التعامل مع نص خام. ولذلك هناك مرحلة تسبق مرحلة تنفيذ خوارزميات تعلم الآلة وهي مرحلة تحويل النص إلى ما يسمى بالميزات Features.

فمثلاً يمكن أن نعبّر عن نص الإيميل بميزاته، مثل عدد الكلمات، عدد الجمل، تواتر وجود كلمات مفتاحية محددة، إلخ. نلاحظ الآن في هذه الحالة أننا نتعامل مع الإيميل كشعاع من الميزات feature vector وهذا أمر مناسب جداً للعديد من خوارزميات تعلم الآلة. أيضاً إن الميزات التي ذكرناها هي ميزات عددية numerical features، ولكن بشكل عام يمكن أن تكون الميزات هي ميزات نصية string features أو ميزات صنفية categorical features، إلخ. ويمكن أيضاً تنميط الميزات بشكل مختلف أو أكثر دقة مثل تصنيف الميزات

العددية إلى ميزات مستمرة continuous features وميزات متقطعة discrete features. وتعود طريقة التنميط إلى التطبيق أو خوارزميات تعلم الآلة المستخدمة. تسمى هذه المرحلة بمرحلة استخراج الميزات Feature Extraction.

في التطبيقات الواقعية تسبق المرحلة السابقة مرحلتين أساسيتين. مرحلة تجميع المعطيات، ومرحلة تنظيفها. تتم عملية تجميع المعطيات بحسب التطبيق. فمثلاً قد تكون المعطيات هي نتيجة استبيانات، أو إحصائيات، أو تم الحصول عليها من مواقع إلكترونية، إلخ. مرحلة تنظيف المعطيات تهدف إلى التأكد سلامة المعطيات قبل استخدامها. وقد تتم هذه العملية بشكل يدوي أو بشكل مؤتمت وذلك بحسب مصدر المعطيات ونظافتها.

3.1.2 خوارزميات تعلم الآلة

كما رأينا في الفقرة 1.1.2، هناك العديد من أصناف المسائل الممكن حلها باستخدام تعلم الآلة. تصنف خوارزميات تعلم الآلة تبعاً لصنف المسألة التي تقوم بحلها. فمثلاً يمكن استخدام الانحدار الخطي Linear Regression لحل مسائل الانحدار [25]. أو استخدام خوارزمية K-Means Clustering لحل مسائل التجميع [25]. سنمهد في هذه الفقرة لأهم خوارزمية مستخدمة في هذا المشروع. وهي الـ SVM.

خوارزمية الـ SVM

إن كلمة SVM هي اختصار لـ Support Vector Machine. وهي خوارزمية تصنيف شهيرة وواسعة الاستخدام في تطبيقات تعلم الآلة. تعتبر خوارزمية قوية حيث أنها تستند على أساس رياضي متين، ولها عدد من الخصائص المهمّة. يمكن تقديم هذه الخوارزمية بعدّة طرق. سنقدمها بطرح مسألة الأملثة التي تقوم بحلها.

بدايةً لنفرض أن مسألتنا هي مسألة تصنيف وعدد الصفوف هو 2. نرمز بـ $(x^{(i)}, y^{(i)})_{1\leq i\leq m}$ إلى معطيات التدريب. حيث m عدد معطيات التدريب. ويكون المثال التدريبي رقم i، له الصف $y^{(i)}$. مع كون $y^{(i)} = +1$ في حال الصف الأول، و $y^{(i)} = -1$ في حال الصف الثاني. وإن $x^{(i)}$ هو شعاع عددي بـ $n+1$ بُعد أي $x^{(i)} \in \mathbb{R}^{n+1}$. وهو ما سميناه شعاع الميزات في الفقرة 2.1.2. أي هنا لدينا n ميزة، حيث لتبسيط العلاقات الرياضية نضيف $x_0^{(i)} = 1$.

النموذج المطروح في خوارزمية الـ SVM، هو تعريف تابع $f : \mathbb{R}^{n+1} \to \{-1, +1\}$. حيث أننا نقول أنه لأجل عيّنة ما (x, y)، فإنها تنتمي إلى الصف الأول في حال كان $f(x) \geq 0$، وتنتمي إلى الصف الثاني في

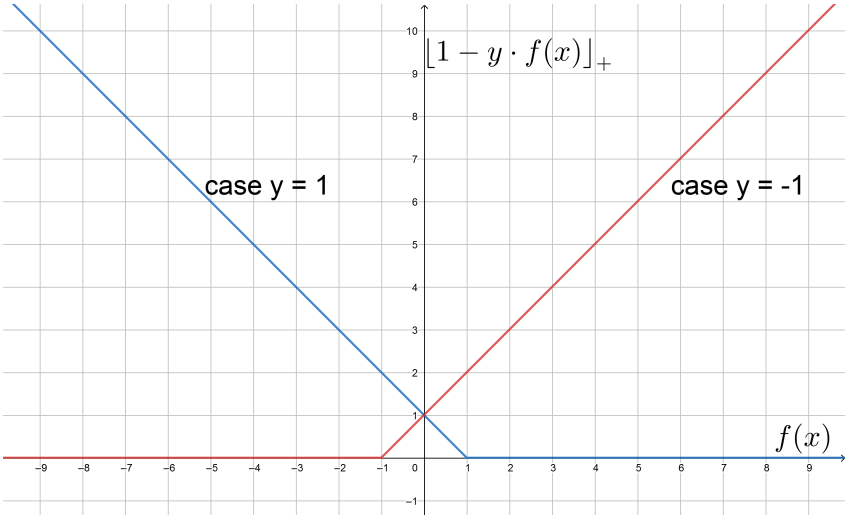
حال $f(x) < 0$. سنأخذ مبدئياً للتبسيط التابع f بالشكل $f(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n$ حيث $\theta = (\theta_j)_{0 \le j \le n}$ هي البارامترات التي يمكن تغييرها.

مسألة الأمثلة التي نريد حلها هي:

$$\underset{\theta \in \mathbb{R}^{n+1}}{\operatorname{argmin}} \left(\|\theta\|_2^2 + C \cdot \sum_{i=1}^{m} \left\lfloor 1 - y^{(i)} f(x^{(i)}) \right\rfloor_+ \right) \quad (1)$$
$$\text{where} f(x^{(i)}) = \sum_{j=0}^{n} \theta_j x_j^{(i)}$$

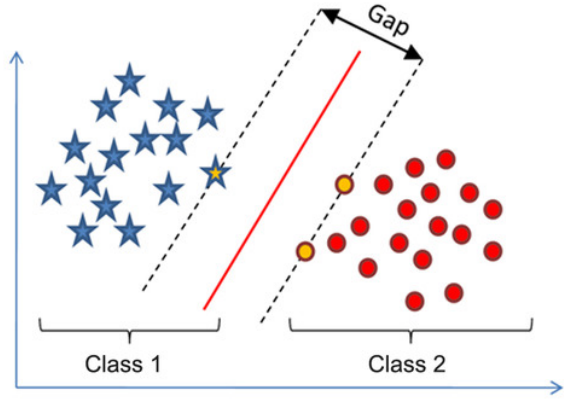
حيث أن التابع $\lfloor \cdot \rfloor_+$ هو تابع الجزء الموجب؛ أي $\lfloor z \rfloor_+ = \max(z, 0)$. و $\|\cdot\|_2$ هو النظيم الإقليدي؛ أي $\|\theta\|_2^2 = \sum_{j=0}^{n} \theta_j^2$. والبارامتر C هو معامل وزن، يحدد مدى التفضيل والمساومة بين الحديّن الأول والثاني في المعادلة. وهو بارامتر فوقي Hyperparameter أي يجب تحديده قبل البدء بحل مسألة الأمثلة، وإن تغييره يغيير حل المسألة.

إن الحد الأول $\|\theta\|_2^2$ في المعادلة 1 هو للتنظيم Regularization. هذا الحد يضبط قيم البارميتر θ ويمنعها من أن تأخذ قيم كبيرة. الحد الثاني يمثل مجموع قيمة الخطأ الحاصل في كل مثال تدريبي من معطيات التدريب. حيث أن الخطأ الحاصل في عيّنة ما (x, y) هو $\lfloor 1 - y \cdot f(x) \rfloor_+$. يمكن تأمل صفات هذا الخطأ من خلال الشكل 1.2. حيث نلاحظ مثلاً في حالة $y = 1$ أن الخطأ يساوي الصفر عندما $f(x) \ge 1$ وأنه يتزايد بشكل خطي كلما أبتعدت قيمة $f(x)$ عن 1 بالاتجاه الخاطئ.



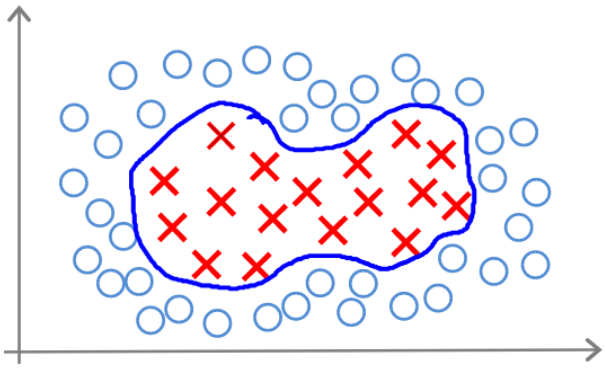
الشكل 1.2: الخطأ في العينة الواحدة في نموذج الـ SVM.

يمكن البرهان على أنه في حالة كون المعطيات قابلة للفصل بخط مستقيم، فإن حل مسألة الأمثلة سيعطي المستقيم f الذي يحقق أكبر هامش ممكن؛ أي اذا قمنا بحساب البعد بين كل نقطة وهذا المستقيم، فإن أصغر بعد سيكون أكبر ما يمكن، وهذا ما يوضحة الشكل 2.2.



الشكل 2.2: مستقيم يفصل صفين بهامش أعظمي.

ولكن أيضاً يمكننا اختيار تابع غير خطي. هذا مفيد مثلاً في حال كان شكل معطيات التدريب مثلما في الشكل 3.2. إذ يوجد أسلوب يسمى بالـ Kernel Trick، يسمح لنا بفعل هذا. ينص هذا الأسلوب على



الشكل 3.2: معطيات التدريب غير قابلة للفصل باستخدام مستقيم.

تعريف f بالشكل $f(x) = \sum_{i=1}^{m} \theta_i K(x, x^{(i)}) + \theta_0$، ثمّ حل مسألة الأمثلة السابقة ذاتها. نلاحظ هنا أنه لدينا $m+1$ باراميتر عوض الـ $n+1$ باراميتر في الحالة السابقة. و يسمى التابع K بالنواة Kernel. فمثلاً اختيار $K(u, v) = \sum_{j=1}^{n} u_j v_j$ يؤدي إلى حل مكافئ لحل المسألة الموضحة في المعادلة 1. تسمى هذه النواة بالنواة الخطية Linear Kernel.

إنّ أشهر النوى المستخدمة عادةً هي:

النواة الخطية	Linear Kernel	$K(u,v) = u^T v$
النواة الحدودية	Polynomial Kernel	$K(u,v) = (u^T v + r)^d$
النواة الغاوسية	Gaussian Kernel	$K(u,v) = \exp\left(-\gamma \left\|u - v\right\|_2^2\right)$

إذ أن الرمز u^T يرمز إلى المنقول وتحديداً فإن $u^T = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}^T = (u_1, \ldots, u_n)$. تسمى النواة الغاوسية أيضاً بـ Radial Basis Function (RBF) Kernel. ونوه أن البارمترات المذكورة r, d, γ هي بارامترات فوقية.

لحل مسألة الأمثلة المطروحة، توجد العديد من الخوارزميات. هذا النوع من المسائل، ومسائل الأمثلة بشكل عام هو فرع مدروس بشكل جيد في الرياضيات تحت اسم Mathematical Optimization. فتوجد العديد من الخوارزميات المستخدمة لحل مسألة الأمثلة المطروحة. من أشهرها هي خوارزمية Sequential Minimal Optimization (SMO). يتطلب شرحها الخوض في كثير من التفاصيل الرياضية وهو خارج نطاق هذا المشروع.

آخر ما يجب ذكره، هو الأسلوب المستخدم للتصنيف في حال وجود أكثر من صفين. هذا الأسلوب مستخدم بشكل عام ويسمى بـ One-versus-All multi-class classification. الفكرة كالتالي، لأجل كل صف، نعتبر جميع الصفوف الأخرى هي صف آخر. ونبني لكل صف، مُصنّف على هذا الأساس. الآن لأجل دخل جديد، نحسب f لكل مُصنّف، ونأخذ الصف الذي يحقق أكبر قيمة.

4.1.2 معايير التقييم

تختلف معايير تقييم صحة نماذج تعلم الآلة باختلاف نوع المسائل التي تقوم بحلها. سنتحدث في هذه الفقرة عن أهم معايير التقييم المستخدمة في مسائل التصنيف.

بدايةً لنضع بعض الرموز لتبسيط العلاقات الرياضية وتوضيح الأفكار. كما تحدثنا سابقاً عن معطيات التدريب، من المعتاد أن توجد معطيات أخرى مستقلة عن معطيات التدريب تسمى بمعطيات الاختبار Test Set. حيث أنه بعد الحصول على النموذج الناتج من خوارزمية تعلم الآلة بتدريبه على معطيات التدريب، يتم اختبار هذا النموذج على معطيات الاختبار. سنرمز لها بـ TS. سنرمز لمجموعة عناصرها بـ (x_i, y_i)، حيث x_i هو شعاع الميزات، y_i هو الصف الموافق. وسنرمز بـ $\hat{y}_i$ للصف الذي تنبأت به خوارزمية تعلم الآلة المستخدمة والتي نريد تقييمها. وسنستخدم الرمز $|\cdot|$ لعدد عناصر مجموعة ما. فمثلاً إن $|y_i = c|$ هو عدد العناصر من TS التي لها الصف c.

الصِحّة Accuracy هي المعيار الأشهر. فهي نسبة العينات التي تم تصنيفها بشكل صحيح. أي:

$$\text{Accuracy} = \frac{|\hat{y}_i = y_i|}{|\text{TS}|}$$

إنّ هذا المعيار غير كافي للتعبير عن مدى قوة النموذج الناتج. لنتأمل مثال تكون فيه معطيات التدريب فيها صفين فقط. نسبة ورود الصف الأول هو 1%، مثل حالة تشخيص مرض نادر. فبإمكاننا بسهولة الحصول على نموذج بدقة 99%. هذا النموذج يتنبأ دائماً بالصف الثاني؛ فلكون ورود عينات تنتمي للصف الأول نادر جداً تكون صحة هذا النموذج عالية. ولكن من الواضح أن هذا النموذج غير مجدي. النقاش السابق يدفع لتحديد معايير أخرى للتقيم.

الدقّة Precision هي معيار يعبر عن دقّة تصنيف صف معيّن. دقّة تصنيف الصف c هي نسبة العينات التي صنفت بشكل صحيح في الصف c من بين جميع العينات التي صنفت بالصف c. أي:

$$\text{Precision for class c} = \frac{|\hat{y}_i = c \wedge y_i = c|}{|\hat{y}_i = c|}$$

الإرجاع Recall هو معيار يعبر عن مدى استرجاعنا لعينات من صف معيّن. معيار الإرجاع للصف c هو نسبة العينات التي صنفت بشكل صحيح في الصف c من بين جميع العينات التي هي ضمن الصف c فعلاً. أي:

$$\text{Recall for class c} = \frac{|\hat{y}_i = c \wedge y_i = c|}{|y_i = c|}$$

المعيار الأخير الذي سنتحدث عنه يسمى بـ F1-score. ينتج من حاجتنا إلى الاعتماد على قيمة عددية واحدة فقط لمقارنة نموذجين معاً. وهو معيار يجمع بين الدقّة والإرجاع. النموذج المقترح للجمع بينهما هو:

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

سنرمز لهذا المعيار اختصاراً بـ F-score. حيث أن الرقم 1 في اسمه يدل على أننا نعطي للدقّة والإرجاع نفس الأهمية. فهذا المعيار حالة خاصة من معيار أعم يسمح بإعطاء أهمية أكبر للدقّة على الإرجاع وبالعكس، ولكن لن نتحدث عنه.

2.2 معالجة اللغات الطبيعية

معالجة اللغات الطبيعية Natural Language Processing (NLP) هو المجال الذي يدرس الآليات التي تسمح للحواسيب والآلات بفهم ومعالجة اللغات الطبيعية مثل اللغة العربية والإنكليزية وغيرهما [26]. ويعتبر مجال مهم في الذكاء الصنعي. يتقاطع هذا المجال مع العديد من المجالات منها علم اللسانيات وتعلم الآلة وغيرهما. سنتحدث في هذه الفقرة بشكل بسيط عن أهم المراحل في معالجة اللغات الطبيعية.

- التحليل الصرفي Morphological Analysis وهو المرحلة التي يجري فيها تحليل الكلمة إلى مكوناتها الأساسية. يُنفّذ هذا التحليل على مستوى الكلمة دون النظر إلى السياق. بشكل أساسي هناك مرحليتن لهذا التحليل. التقطيع Tokenization والتشذيب Stemming.

  خرج مرحلة التقطيع هو الرموز التي تكون الجملة Tokens. أي مثلاً إنّ الجملة

  Google inc. is huge.

  سيتم تقسيمها إلى خمس رموز وهي:

  {Google | inc. | is | huge | .}

  لاحظ أن النقطة الأخيرة تعتبر رمز منفصل بينما النقطة في الكلمة inc. ليست رمز منفصل.

  خرج مرحلة التشذيب هو جذر الكلمة، بالإضافة إلى السوابق واللواحق، ومعلومات أخرى تختلف باختلاف اللغة. مثلاً إن جذر الكلمات fish, fishing, fished, fisher هو fish. واللواحق هي ing, ed, er على الترتيب.

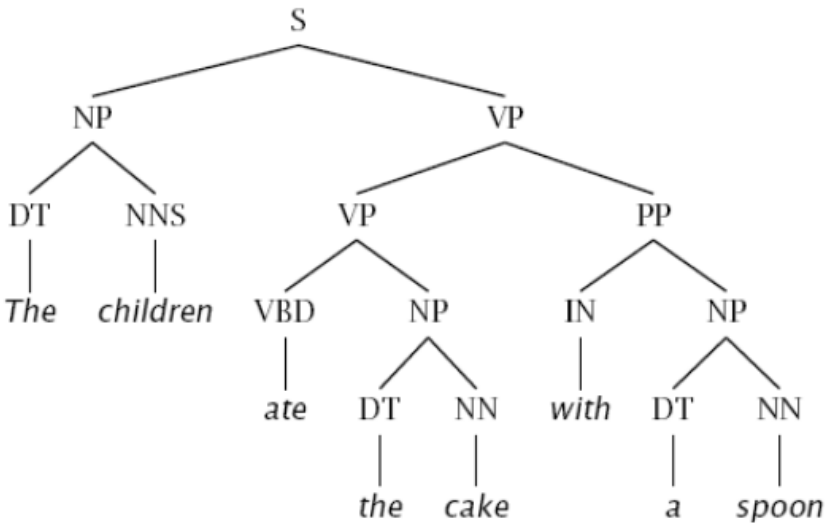
- تحديد أنماط الكلمات Part-of-Speech Tagging وهو عملية إسناد الأنماط النحوية الملائمة لكل كلمة من كلمات الجملة. دخل هذه المرحلة عادةً يكون كلمة ضمن سياق محدد (ضمن جملة). الخرج الناتج يكون النمط النحوي لهذه الكلمة (اسم، فعل، صفة، إلخ).

- التكتيل Chunking وهو تقسيم الجملة إلى عبارات أصغر؛ أي عبارات اسمية، أو عبارات فعلية، إلخ. يوضح الشكل 4.2 مثال على ذلك.

[NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

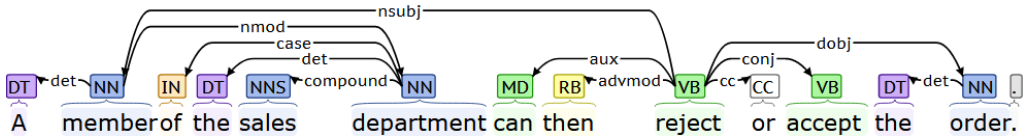
الشكل 4.2: عمليّة التكتيل للجملة He reckons the current account deficit will narrow to only # 1.8 billion in September. تم تقسيم هذه الجملة إلى جمل فعلية Verb Phrase (VP)، وجمل اسمية Noun Phrase (NP)، وعبارات جر Prepositional Phrase (PP).

- الشجرة النحوية Parsing Tree. في الواقع يمكن اعتبار الشجرة النحوية الخرج الناتج عن العمليات السابقة مجتمعة. حيث يتم تمثيل الجملة كشجرة. انظر الشكل 5.2 لمثال توضيحي. الاختصارات المكتوبة، مثل NN تعني اسم مفرد singular noun، NNS تعني اسم جمع plural noun. وللإطلاع على هذه القائمة كاملةً انظر [1].



الشكل 5.2: الشجرة النحوية الناتجة للجملة The children ate the cake with a spoon.

- تحليل التّبعية Dependency Parsing. نظراً لقصور الشجرة النحوية عن إعطاء كامل المعلومات حول الجملة المدروسة، تم اقتراح تمثيل المعلومات النحوية على شكل بيان التبعية Dependency Graph. يوضح هذا البيان العلاقات النحوية بين كلمات الجملة. إذ يتكون من عقد تمثّل الكلمات ووصلات تحدد العلاقة بين هذه الكلمات. يبيّن الشكل 6.2 مثالاً على خرج تحليل التبعية. الاختصارات المكتوبة للعقد هي ذاتها المستخدمة في حالة الشجرة النحوية. الاختصارات المكتوبة للوصلات، مثل det تربط بين الاسم وأداة التعريف المرتبطة به، nsubj تربط بين الفعل والفاعل. وللإطلاع على هذه القائمة كاملةً انظر [7].



الشكل 6.2: بيان التبعية الناتج للجملة A member of the sales department can then reject or accept the order.

- حل الغموض في حالات الإحالة Anaphora Resolution. الإحالة Anaphora هي استخدام

ضمائر أو تعابير للإشارة إلى اسم أو تعبير تم ذكره في سياق سابق في النص. فإذا تأملنا المثال التالي:

They buy the issue, then resell it to the public.

إن it تشير إلى issue. وفي هذه الحالة يكون المشار إليه واقعاً في نفس الجملة. أمّا في حالة They فيكون المشار إليه واقعاً في جملة سابقة.

3.2 الأوراق العلمية

كما شرحنا في الفقرة 1.2 وتحديداً الفقرة 2.1.2، إن أولى الخطوات اللازمة لتنجيز أي نظام يستخدم تعلم الآلة هي تحديد الميزات التي سيتم استخدامها. فهذا المشروع يتعامل مع النصوص وكون المسألة المطروحة هي تقييم جودة هذه النصوص من ناحية سهولة القراءة، فمن المهم جداً معرفة الميزات التي ستعبّر عن وتوصّف جودة نص.

بالعودة إلى العديد من الأوراق العلمية التي تتقاطع مع هذا المشروع، تم تجميع عدد من الأفكار والمنهجيات التي تم تبنيها والاعتماد عليها كإطار عمل ضمن المشروع. سنذكر في هذه الفقرة أهم الأوراق العلمية التي تمت دراستها، وأهم الأفكار والمنهجيات المتبعة فيها. بالإضافة إلى الميزات وخوارزميات تعلم الآلة المستخدمة لحل مسائل مشابهة للمسألة المطروحة في هذا المشروع.

إن فكرة تقييم النصوص من ناحية صعوبة القراءة بشكل موضوعي (غير شخصي) بدأت تقريباً منذ قرن. الأفكار الأولية التي واجهت هذا الموضوع هي عبارة عن علاقات رياضية بسيطة. ولقد اعتَمَدَت على خصائص سطحية في النص المدروس. مثل متوسط طول الجملة، ومتوسط طول الكلمة. فمثلاً إن Flesch Score يعطى بالعلاقة

$$\text{Flesch Score} = 206.835 - 1.015 \cdot \frac{\#\text{words}}{\#\text{sentences}} - 84.6 \cdot \frac{\#\text{syllables}}{\#\text{words}}$$

أي يمثل علاقة خطية لمتوسط طول الجملة بالكلمات، ومتوسط طول الكلمة بالمقاطع الصوتية. وكلما كان هذا المقدار أكبر، كلما كان النص أسهل للقراءة. في [5] أُجريب مقارنة بين عدد واسع من هذه العلاقات. وفي [6] تم رسم ودراسة توزع عدد من هذه المعايير على عدد كبير من النصوص. على الرغم من كون هذه العلاقات تبدو سطحية من ناحية التمثيل اللغوي للنص، تم اعتمادها بشكل واسع لمدة من الزمن.

كما ظهرت نماذج أكثر تعقيداً تعتمد على مفهوم الـ n-gram. وهو نموذج إحصائي لتمثيل اللغة؛ يعبر عن احتمال ورود كلمة ضمن سياق معيّن، أو احتمال ورود جملة أو سياق معيّن. وهو تمثيل بسيط للانتروبية في

اللغة الانكليزية. استخدم هذا المفهوم بالإضافة إلى ميزات أخرى بُنيت فوقه ومستوحاة من مفهوم الانتروبية في عدد من الدراسات أهمها [9,3]. وكانت النتائج أفضل بشكل واضح عن نتائج المعايير السابقة والمعتمدة على علاقات رياضية بسيطة. حيث تم الاعتماد على خوارزميات تعلم الآلة. الخوارزمية الأكثر استخداماً والتي حققت أفضل النتائج كانت الـ SVM.

مع تطور الأدوات في معالجة اللغات الطبيعية، أصبح من الممكن أن نأخذ بعين الاعتبار مؤشرات أقوى، مفرداتية ونحوية وغيرها. في [11,10] تمت دراسة ومقارنة العديد من الميزات التي تعتمد بشكل أساسي على تقدم الأدوات في معالجة اللغات الطبيعية. حيث تمت مقارنة عدّة ميزات، وبعدّة أنواع. مثل التنوع في الأفعال، والكلمات المستخدمة، وكثافة الأسماء المستخدمة في النص. بالإضافة إلى الترابط بين الجمل باستخدام بيان التبعية والعديد غيرها. أفضل نتيجة تم تحقيقها كانت باستخدام الـ SVM على مجموعة جزئية من الميزات المدروسة.

أيضاً لقد تمت دراسة تعقيد النصوص المكتوبة من قبل الطلاب الذين يتعلمون اللغة الانكليزية. تم ذلك تحت ما يسمى أبحاث تعلم اللغات second language acquisition. حيث تمت دراسة عدد من المؤشرات التي تعبر عن تعقيد النص، بما يفيد في دراسة تحسن الطلاب أثناء تعلمهم للغة. كما أن دراسات لاحقة قامت بأتمتت آليات حساب هذه المؤشرات أهمها [12]. فكان تغيير هذه المؤشرات مع الزمن، يبيّن مستوى تحسن الطلاب في تعلم اللغة. وكان أول استخدام لهذه الدراسات لبناء نموذج تعلم آلة لتقييم جودة النصوص هو في [15]. حيث تم الاعتماد على ميزات مستوحاة بشكل مباشر من هذه المؤشرات. معظم هذه الميزات تعبر عن تعقيد تراكيب الجمل. مثل وجود جمل شرطية أو جمل معطوفة على بعضها وهكذا.

أيضاً تم استخدام ميزات تعبر عن نسبة ورود كلمات معينة ضمن النص. فإن ورود كلمات متقدمة ضمن النص وبتواتر عالي قد يكون مؤشر جيد على كون النص بمستوى متقدم. هذه الكلمات تكون غالباً منتقاة من مصدر تعليمي. فقد تم استخدام كلمات من المصدر The Academic Word List[1] في [27,18,15]. و تم استخدام كلمات من المصدر English Vocabulary Profile[2] في [24]. حيث تبيّن أن لهذه الميزات دور جيد في تحسين أداء النماذج الناتجة.

سنوضح لاحقاً وبتفصيل أكبر في الفقرة **??** الميزات المستخدمة في هذا المشروع. يجب أيضاً التنويه إلى المعطيات المستخدمة في هذه الدراسات. إن معظم المعطيات هي من مصدر تعليمي، مثل مجلات تعليمية للأطفال حيث أن المقالات مصنفة بحسب الفئة العمرية المناسبة. سنتحدث لاحقاً في الفقرة **??** عن مجموعة المعطيات المستخدمة في هذا المشروع وخصائصها.

[1] للإطلاع على هذه القائمة كاملةً انظر [2].
[2] لمزيد من التفاصيل انظر http://www.englishprofile.org.

المراجع

[1] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank". In: *Computational linguistics* 19.2 (1993), pp. 313–330.

[2] Averil Coxhead. "A new academic word list". In: *TESOL quarterly* 34.2 (2000), pp. 213–238.

[3] Sarah E Schwarm and Mari Ostendorf. "Reading level assessment using support vector machines and statistical language models". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2005, pp. 523–530.

[4] Rong Zheng et al. "A framework for authorship identification of online messages: Writing-style features and classification techniques". In: *Journal of the American society for information science and technology* 57.3 (2006), pp. 378–393.

[5] William H DuBay. "The Classic Readability Studies." In: *Online Submission* (2007).

[6] Ronald P Reck and Ruth A Reck. "Generating and rendering readability scores for Project Gutenberg texts". In: *Proceedings of the Corpus Linguistics Conference*. 2007.

[7] Marie-Catherine De Marneffe and Christopher D Manning. *Stanford typed dependencies manual*. Tech. rep. Technical report, Stanford University, 2008.

[8] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. "Cognitively motivated features for readability assessment". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2009, pp. 229–237.

[9] Sarah E Petersen and Mari Ostendorf. "A machine learning approach to reading level assessment". In: *Computer speech & language* 23.1 (2009), pp. 89–106.

[10] Lijun Feng. "Automatic readability assessment". In: (2010).

[11] Lijun Feng et al. "A comparison of features for automatic readability assessment". In: *Proceedings of the 23rd international conference on computational linguistics: Posters*. Association for Computational Linguistics. 2010, pp. 276–284.

[12] Xiaofei Lu. "Automatic analysis of syntactic complexity in second language writing". In: *International journal of corpus linguistics* 15.4 (2010), pp. 474–496.

[13] Scott A Crossley, David B Allen, and Danielle S McNamara. "Text readability and intuitive simplification: A comparison of readability formulas." In: *Reading in a foreign language* 23.1 (2011), pp. 84–101.

[14] Scott A Crossley, David Allen, and Danielle S McNamara. "Text simplification and comprehensible input: A case for an intuitive approach". In: *Language Teaching Research* 16.1 (2012), pp. 89–108.

[15] Sowmya Vajjala and Detmar Meurers. "On improving the accuracy of readability classification using insights from second language acquisition". In: *Proceedings of the seventh workshop on building educational applications using NLP*. Association for Computational Linguistics. 2012, pp. 163–173.

[16] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[17] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[18] Sowmya Vajjala and Detmar Meurers. "Readability assessment for text simplification: From analysing documents to identifying sentential simplifications". In: *ITL-International Journal of Applied Linguistics* 165.2 (2014), pp. 194–222.

[19] Johann Schleier-Smith. "An architecture for Agile machine learning in real-time applications". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2015, pp. 2059–2068.

[20] D Sculley et al. "Hidden technical debt in machine learning systems". In: *Advances in neural information processing systems*. 2015, pp. 2503–2511.

[21] Sowmya Vajjala. "Analyzing text complexity and text simplification: connecting linguistics, processing and educational applications". PhD thesis. Ph. D. thesis, University of Tübingen, 2015.

[22] Sowmya Vajjala and Detmar Meurers. "Readability-based sentence ranking for evaluating text simplification". In: *arXiv preprint arXiv:1603.06009* (2016).

[23] Ian H Witten et al. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[24] Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. "Text readability assessment for second language learners". In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. 2016, pp. 12–22.

[25] Aurélien Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2017.

[26] Riad Sonbol. *Extracting Business Process Models from Natural Language Texts*. Higher Institute for Applied Sciences and Technology, 2017.

[27] Sowmya Vajjala and Ivana Lucic. “OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification”. In: (2018).