

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

تقييم جودة نصوص اللغة الانكليزية

مشروع السنة الرابعة

إعداد

فاروق حجابو

إشراف

م. رياض سنبل

د. غيداء ريداوي

2 آب 2018

الملخص

المحتويات

i	الغلاف
ii	الملخص
iii	المحتويات
v	قائمة الأشكال
vi	قائمة الجداول
vii	الاختصارات
ix	المصطلحات
1	1 التعريف بالمشروع
1	1.1 مقدمة
2	2 الدراسة المرجعية
2	1.2 تعلم الآلة
3	1.1.2 تصنيفات تعلم الآلة

4	المراحل اللازمة لتطبيق تعلم الآلة	2.1.2
5	خوارزميات تعلم الآلة	3.1.2
8	معايير التقييم	4.1.2

قائمة الأشكال

1.2	الخطأ في العينة الواحدة في نموذج الـ SVM	6
2.2	مستقيم يفصل صفين بهامش أعظمي	7
3.2	معطيات التدريب غير قابلة للفصل باستخدام مستقيم	7

قائمة الجداول

الاختصارات

SVM Support Vector Machine

SMO Sequential Minimal Optimization

المصطلحات

Artificial Intelligence	الذكاء الصناعي
Machine Learning	تعلم 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	الصحة
Precision	الدقة
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	بارامتر فوق

الفصل الأول

التعريف بالمشروع

يُمهّد هذا الفصل للمشروع، حيث يُبيّن فكرة المشروع وأهميتها والأهداف المرجوة 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. ويقدم 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 الاصطناعي Artificial Intelligence. يُقصد بتعلم الآلة مجموعة الأدوات والمفاهيم والمنهجيات المستخدمة لبرمجة الحواسيب بطريقة تسمح لهذه الحواسيب بالتعلم من المعطيات.

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 بفلتر 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 بعدة طرق؛ فمثلاً يمكننا استخدام نسبة عدد الإيميلات التي تم تصنيفها بشكل صحيح إلى عدد الإيميلات الكلي (هذا المعيار يسمى الصحة Accuracy كم سنرى لاحقاً).

1.1.2 تصنيفات تعلم الآلة

يمكن تصنيف أنظمة تعلم الآلة وفق عدة 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 معين. على سبيل المثال، قد ترغب شركة في تصنيف زبائنهم إلى عدة 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 Semi-Supervised Learning: وهي حالة وسيطة بين التصنيفين السابقين. تكون فيها بعض أمثلة التدريب مرتبطة بخرج معين (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 معين من معطيات التدريب) هو إمّا مؤذّي أو غير مؤذّي.
- 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 نعبر عن نص الإيميل بمميزاته، 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 تمثيل الميزات بشكل مختلف أو أكثر دقة مثل تصنيف الميزات

العددية إلى ميزات مستمرة continuous features وميزات متقطعة discrete features. وتعود طريقة الترميز 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 [11]. أو استخدام خوارزمية K-Means Clustering لحل مسائل التجميع [11]. سنمهد في هذه الفقرة لأهم خوارزمية مستخدمة في هذا المشروع. وهي ال SVM.

خوارزمية ال SVM

إن كلمة SVM هي اختصار ل 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 المهمة. يمكن تقديم هذه الخوارزمية بعدة طرق. سنقدمها بطرح مسألة الأمثلة 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 ال SVM، هو تعريف تابع $f: \mathbb{R}^{n+1} \rightarrow \{-1, +1\}$. حيث أننا نقول أنه لأجل عينة 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 $f(x) = \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_n x_n$ حيث $\theta = (\theta_j)_{0 \leq j \leq n}$ هي البارامترات التي يمكن تغييرها.

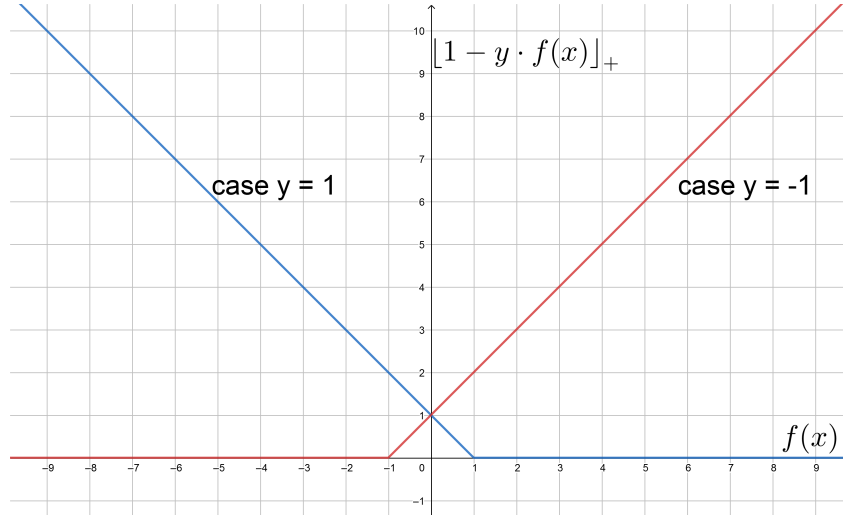
مسألة الأمثلة التي نريد حلها هي:

$$\operatorname{argmin}_{\theta \in \mathbb{R}^{n+1}} \left(\|\theta\|_2^2 + C \cdot \sum_{i=1}^m [1 - y^{(i)} f(x^{(i)})]_+ \right) \quad (1)$$

where $f(x^{(i)}) = \sum_{j=0}^n \theta_j x_j^{(i)}$

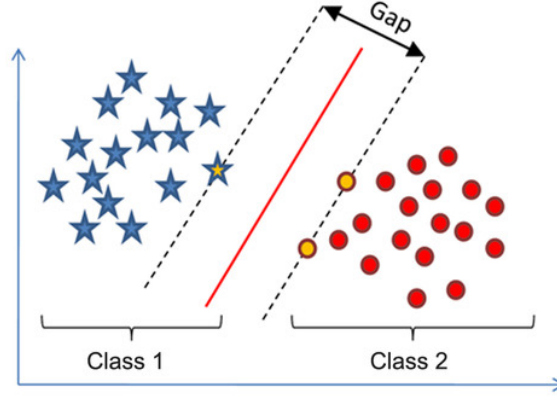
حيث أن التابع $[\cdot]_+ = \max(z, 0)$ هو تابع الجزء الموجب؛ أي $\|z\|_+ = \max(z, 0)$ و $\|\cdot\|_2$ هو التنظيم الإقليدي؛ أي $\|\theta\|_2^2 = \sum_{j=0}^n \theta_j^2$. والبارامتر C هو معامل وزن، يحدد مدى التفضيل والمساومة بين الحدين الأول والثاني في المعادلة. وهو بارامتر فوق Hyperparameter أي يجب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

إن الحد الأول $\|\theta\|_2^2$ في المعادلة 1 هو للتنظيم Regularization. هذا الحد يضبط قيم البارامتر 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 عينة ما (x, y) هو $[1 - y \cdot f(x)]_+$. يمكن تأمل صفات هذا الخطأ من خلال الشكل 1.2. حيث نلاحظ مثلاً في حالة $y = 1$ أن الخطأ يساوي الصفر عندما $f(x) \geq 1$ وأنه يتزايد بشكل خطي كلما أبتعدت قيمة $f(x)$ عن 1 بالاتجاه الخاطئ.



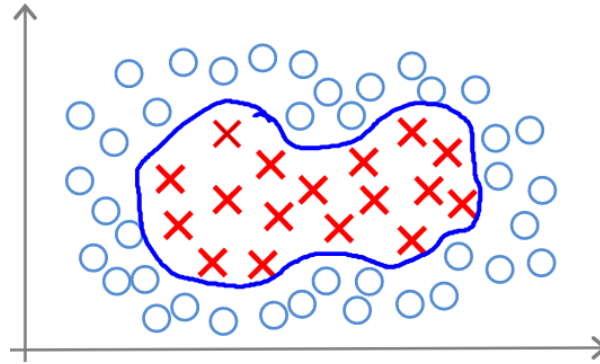
الشكل 1.2: الخطأ في العينة الواحدة في نموذج الـ SVM.

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 إذا قمنا بحساب البعد بين كل نقطة وهذا المستقيم، فإن أصغر بعد سيكون أكبر ما يمكن، وهذا ما يوضحه الشكل 2.2.



الشكل 2.2: مستقيم يفصل صفتين 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 بال Kernel Trick، يسمح لنا بفعل هذا. ينص هذا الأسلوب على



الشكل 3.2: معطيات التدريب غير قابلة للفصل باستخدام مستقيم.

تعريف f بالشكل $f(x) = \sum_{i=1}^m \theta_i K(x, x^{(i)}) + \theta_0$ ، ثم حل مسألة الأمثلة السابقة ذاتها. نلاحظ هنا أنه لدينا $m + 1$ بارامتر عوض الـ $n + 1$ بارامتر 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.

إن أشهر النوى المستخدمة عادةً هي:

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

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

حل 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.

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 .

الصحة 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 كافٍ للتعبير عن مدى قوة النموذج الناتج. لتأمل مثال تكون فيه معطيات التدريب فيها صنفين 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 للتقييم.

الدقة Precision هي معيار يعبر عن دقة تصنيف صف معيّن. دقة 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 الدقة والإرجاع. النموذج المقترح للجمع بينهما هو:

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

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

المراجع

- [1] 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
- [2] 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
- [3] 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
- [4] Sarah E Petersen and Mari Ostendorf. “A machine learning approach to reading level assessment”. In: Computer speech & language 23.1 ,(2009) pp. –89.106
- [5] Xiaofei Lu. “Automatic analysis of syntactic complexity in second language writing”. In: International journal of corpus linguistics 15.4 ,(2010) pp. –474.496

- [6] 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
- [7] 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
- [8] Sowmya Vajjala. “Analyzing text complexity and text simplification: connecting linguistics, processing and educational applications”. PhD thesis. Ph. D. thesis, University of Tübingen, .2015
- [9] Sowmya Vajjala and Detmar Meurers. “Readability–based sentence ranking for evaluating text simplification”. In: arXiv preprint arXiv:1603.06009 ,(2016)
- [10] 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
- [11] Aurélien Géron. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. ” O’Reilly Media, Inc.”, .2017
- [12] Riad Sonbol. Extracting Business Process Models from Natural Language Texts. Higher Institute for Applied Sciences and Technology, .2017
- [13] Sowmya Vajjala and Ivana Lucic. “OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification”. In: .(2018)