جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصنعي

AUTOMATED ESSAY SCORING

[التقييم الآلي للمقالات باللغة الإنكليزية]

مشروع أعد لنيل درجة الإجازة في الهندسة المعلوماتية

إعداد الطلاب:

حلا عويس وسام السحلي

عبير السيد تسنيم زغموت

بإشراف:

د. ندى غنيم

م. علا طبال

2022-2021

الإهداء: حلا

إلى من أحمل اسمه بكل فخر... من تشققت يداه في سبيل رعايتي وتحقيق أحلامي... الذي بفضله كنت اليوم هنا

إلى والدي ومعلمي

إلى ملاكي، إلى التي في لمساتها يكمن الحب والحنان، من وهبتني الحياة وصبرت عليّ سنين دراستي، من كان في دعائها سبباً لنجاحي

إلى أمي وحبيبتي

إلى من كان الداعم الأول لي في مسيرتي الدراسية، من كان سنداً ومصدراً للقوة رغم البعد، أكثر من تمنيت وجوده اليوم ليحصد معي فرحة نجاحي

إلى أخي وعزيزي عامر

إلى أحب الناس لقلبي وأغلاهن من لا تحلو الأيام إلا بوجودهن، من أنرن لي الطريق، وشددن على يدي لحظات ضعفي، من وجودهنّ يدفعني لأقدم أفضل ما لديّ، من يحافظن على أحلامي أكثر مني ويساعدنني على تحقيقها.

الغاليات أخواتي رنيم، مرح، آية، لين

إلى من هوّنوا عليّ مشقّة الدرب، ساعدوني على تخطي مصاعب الجامعة وتجاوزها بسلاسة، من وقفوا بجانبي دوماً وصفقوا لنجاحي، من تقاسمنا الضحكات والذكريات ،إلى أثمن من وهبتني الجامعة

ليلى، بتول، دعاء، تسنيم، براءة، رنيم

إلى من بمساعدتهم تمكنت من الاستمرار، من شاركوني أعمالي الجامعية وصعوباتها والضغوطات، وقاسموني لحظات التعب بكل حب، إلى أصحاب القلوب الطيبة والنوايا الصادقة

نوال، علا، رغد

إلى شريكتي اللطيفة، من شاركتني مصاعب السنة الأخيرة وأيامها، صاحبة القلب الطيب

تسنيم

إلى أصدقاء مقاعد الدراسة والذكريات البريئة، من مهما أبعدتنا الظروف لهم في قلبي ذات المودة إن لم تزد، من دعموا خطواتي الأولى وانتظروا معي نجاحي وفخروا به بكل حب

نور، آية

إلى من شاركوني هذا العمل من تشاركنا صعوباته، ومن عملوا بجد حتى ننجزه

وسام، عبير، تسنيم

شكراً لكل من كان في الطريق، استندت عليه أو مرّ عابراً، شكراً للأهل للرفاق والأصدقاء الذين كانوا المعنى للأيام المكررة، شكراً لكل من علّمني كلمة أو مهارة، شكراً لمن كان سبباً بخلق أي شعور، لم يكن سهلاً لكن لطف الله كان مصاحباً لي

الإهداء.. وسام..

بِسْمِ اللَّهِ الرَّحْمَـٰنِ الرَّحِيمِ

﴿وَقُلِ اعْمَلُوا فَسَيَرَى اللَّهُ عَمَلَكُمْ وَرَسُولُهُ وَالْمُؤْمِنُون﴾

﴿قُلْ بِفَضْلِ اللَّهِ وَبِرَحْمَتِهِ فَبِذَٰلِكَ فَلْيَفْرَحُوا هُوَ خَيْرٌ مِّمَّا يَجْمَعُون﴾

صدَقَ اللَّهُ العَظِيم

ماكنتُ لأفعل لولا أنَّ الله مكنَّني

الحمد للّه ماخُتم جهد ولا تمَّ سعي إلاَّ بفضله

إلهي لا يطيب الليل إلا بشكرك ولا يطيب النهار إلا بطاعتك.. ولا تطيب اللحظات إلا بذكرك.. ولا تطيب الآخرة إلا بعفوك.. ولا تطيب الجنة إلا برؤيتك.

اللَّهُ جلَّ جلالُه...

إلى من بلَّغ الرسالة.. وأدَّى الأمانة.. ونصحَ الأمَّة.. إلى نبيِّ الرَّحمة ونور العالمين..

سيدنا محمَّد صلَّى اللَّهُ عليهِ وسلَّم..

إلى من أرى في عينيها فرحةَ تخرجي.. إلى من كان دعاؤها أمانَ طريقي..

إليكِ أمِّي..

إلى من أحملُ اسمه بكل افتخار.. إلى من شد بيدي وساندني لأصل..

إليكَ أبَي..

إلى زينة منزلنا وفرحة أيامي.. إلى من هم هدية من الرحمن..

إخوَتي: بشَّار وأسامة وريماس..

إلى من وثقوا بي لأصل.. إلى من ألجأ إليهم لحظة ضعفي.. إلى خالاتي:

منال وماجدة وهيام أحمد ..

إلى منبع البدايات الجميلة.. إلى الذي زرع فينا حب العطاء والتعاون ..

إلى فريق الكريات الحمراء..

إلى من كانت خيرَ عونٍ لي أيَّام التعب.. إلى من ساندتني لحظات الضياع..

المهندسة عُلا طبَّال..

إلى نبعَ العطاء والإخلاص في العمل.. إلى من أحبَبت العلمَ لأجلها..

المهندسة زينة الدَّلال..

إلى من ابتدأت رحلتي معهم وقضينا معاً أجملَ اللحظات.. إلى من هم في دعائي دوما..ً

رهام قهوجي وأريـج درويش ورغدة الصّمل وأريـج الدرزي..

إلى أختِ لم تلدها أمِّي.. مصدر تفاؤلي.. إلى من معها تخطيت صِعابي.. إلى أملي..

المهندسة أمل محمود حجازي..

إلى الكتف الذي يسندني وأميلُ عليه..الضحكة التي تخرج من قلبي.. الراحة التي ألقاها بعد التعب..

إلى جُلّنار نعيم وفداء غزلان وهبةُ الله شقّوف ويُمنى الشَّلق..

إلى رفيقة لم تكن رفيقة جامعة ودراسة بل رفيقة حياة وجنَّة بإذنِ الله..

إلى نور هيثم بيرقدار..

إلى صاحبة الهمّة العالية والقلب الطيب.. من زرعت السعادة في قلبي وشريكة المشروع..

عبير سامي السيد..

إلى معلّماتي ومصدر قوتي..من كانت كلماتهم شفاء لروحي..

هيام الأحمر وفضيلة عرب ونور سلمان وفاطمة الخطيب وآمنة ورؤى..

إلى من جمعتني بهم السنة الأخيرة.. معهم ذقت لذَّة الوصول والنجاح.. شركاء المشروع:

حلا عويس وتسنيم زغموت..

إلى أساتذتي ومعلميّ.. إلى كل من علَّمني وأرشدني..

إلى كلِّ من دعا لي دعوةَ في ظهر الغيب.. إليكم أُهدي بحثي..

م وسام السحلي٨٨

الفهرس:

الملخص.....8

الفصل الأول: مقدمة.....9

دوافع العمل.....10

فكرة المشروع 10

اهداف المشروع.....11

شريحة المستخدمين.....11

الفصل الثاني: الدراسة المرجعية.....12

تطبيقات مشابهة.....13

أبحاث مشابهة 14

الفصل الثالث: الدراسة التحليلية.....23

المتطلبات الوظيفية.....24

المتطلبات غير الوظيفية.....24

حالات الاستخدام.....25

Flow Charts.....26

الفصل الرابع: التجارب والاختبار.....28

نموذج تقيم المقال.....29

نموذج ارتباط الموضوع بالمقال.....32

الفصل الخامس: التحقيق.....36

بنية النظام.....37

الموارد المستخدمه.....37

بيانات التدريب..... 37

معايير الاداء.....38

السمات المستعمله.....43
النتائج والمقارنة 52
الفصل السادس: الخاتمه.....53
الخاتمه والافاق المستقبلية.....55
الفصل السابع: المراجع.....56

الملخص:

نقوم بهذا العمل ببناء منصة تعليمية تعتمد على تعلم الآلة ومعالجة اللغات الطبيعية. المنصة مكونة من نظام يقوم بتقييم مقال الطالب وإعطاءه الدرجة المستحقة بتحليل كتابته واكتشاف الأخطاء.

يقوم النظام بتسهيل عملية وضع العلامات للطلاب ويختصر وقت المعلم.

المهام التي قمنا بمعالجتها بهذا العمل: التحقق من ترابط موضوع النص بالسؤال – إعطاء العلامة للمقال المدخل.

حيث تم إنجاز جميع المتطلبات والمهام الأساسية وتم التعميم من مواضيع الداتا لأي مقال يدخله المستخدم.

نود في المستقبل أن نقوم بتجميع داتا أكبر للحصول على نتائج أفضل للنظام.

الكلمات المفتاحية: Automated essay scoring – asap -essay grading – machine learning – natural language processing

الفصل الأول:
مقدمة

مقدمة:

1.1- دو افع العمل:

تقييم المقالات يستغرق وقتًا طويلاً ومكلف إذا أمضى مدرس جامعي ما معدله 10 دقائق في قراءة مقال كل طالب وإعطاءه العلامة، فإن تقييم جميع مقالات الطلاب البالغ عددها 150 مقال سيتطلب 25 ساعة من التقييم بدون توقف. ليس من المستغرب أن يقلل المعلمون عدد مهام الكتابة التي يكلفونها للطلاب خلال الفصل الدراسي لمقالين أو ثلاثة فقط، مما يحرم الطلاب من فرص ممارسة الكتابة الإضافية. يمكن أن يساعد تعيين المزيد من المعلمين لتقليل أحجام الفصول الدراسية، ولكنه سيؤدي إلى زيادة الرسوم الدراسية والضرائب

مع التقدم في تكنولوجيا الكمبيوتر، أصبحت إمكانية تقييم المقالات باستخدام الكمبيوتر حقيقة حيث يسمح Automatic Essay Scoring (AES) للمعلمين بتعيين درجات للمقالات من خلال تحليل الكمبيوتر. يتم استخدم معالجة اللغة الطبيعية (NLP)، وهو شكل من أشكال الذكاء الاصطناعي يمكّن أجهزة الكمبيوتر من فهم اللغة البشرية والتلاعب بها، لتقييم المقالات التعليمية. يعمل برنامج AES عن طريق استخراج ميزات مثل عدد الكلمات، واختيار المفردات، وعدد الأخطاء، وتباين طول الجملة، لإنشاء نموذج إحصائي لجودة المقال. تسمح مقارنة مقال الطالب بهذا النموذج الإحصائي للنظام بتقدير النتيجة في غضون 20 ثانية أو أقل.

في هذا التقرير، نورد نواة نظام تقييم المقالات الكتابية لمساعدة المعلم على تقييم مقالات طلابه بأسرع وقت.

1.2- فكرة المشروع:

المشروع عبارة عن تطبيق الكتروني موجه للطلاب الذين يرغبون بالتدرب على الكتابة والمعلمين الذين يريدون اختصار وقت تقييم مقالات طلابهم.

حيث يقدم التطبيق للطالب موضوع معين لمناقشته ضمن مقالة يكتبها، ويتم فيما بعد تقييم هذه المقالة على التطبيق باستخدام تقنيات الذكاء الصنعي ومعالجة اللغات الطبيعية من أجل إعطاء الطالب فرصة للتدرب قبل الامتحان ومعرفة مستواه.

1.3- أهداف المشروع:

- مساعدة المتقدمين للامتحان على التحضير وتوفير المال والوقت:

لكي يحصل المتقدم للامتحان على علامة عالية، عليه معرفة الطريقة الصحيحة لكتابة المقالة، بالإضافة إلى التدريب على كتابة الكثير من المقالات وتصحيحها من قبل استاذ مختص لمعرفة المشاكل وتفاديها. نظام التصحيح الآلي سيأخذ دور الاستاذ، حيث من الصعب وجود الاستاذ دائما مع الطالب إلا في حالة الدورات التدريبية والتي تكلف المبالغ الباهظة.

- توفير الوقت على المعلم:

حيث أن الوقت الذي يحتاجه النظام من أجل تصحيح مقالة واحدة صغير جداً بالمقارنة مع الوقت الذي يحتاجه شخص لتصحيح نفس المقالة.

- توفير التكلفة المالية على المعاهد والمدارس:

حيث أن المعهد لن يضطر لدفع رواتب الموظفين المسؤولين عن التصحيح هو فقط سيدفع تكلفة شراء نظام التصحيح الالي..

1.4- شريحة المُستخدمين:

- المدارس.
- المعاهد ومراكز تعليم اللغة الإنكليزية.
- جميع الراغبين بتحسين كتابتهم لمقالات اللغة الإنكليزية.

الفصل الثاني
الدراسة المرجعيَّة

2. الدراسة المرجعية:

2.1 - تطبيقات مشابهة:

The Virtual Writing Tutor:

هذا البرنامج عبارة عن مدقق نحوي للغة الإنجليزية لغير المتحدثين باللغة كلغة أصلية يمكن للكتاب والمدونين وغيرهم من المهنيين استخدامه أيضًا. يستفيد المعلمون أيضًا من برنامج Virtual Writing Tutor، حيث يتم التعامل مع مهمة التحرير والتدقيق اللغوي بواسطة هذا التطبيق.

<u>الميزات:</u>

1. توفر للطلاب ملاحظات وتصحيحات حيث يقوم بالتدقيق النحوي والإملائي و فحص علامات الترقيم المستخدمة وإعادة صياغة الجمل كما ويقوم بفحص المفردات و إثرائها.
2. لدى المستخدمين فكرة عن عدد كلماتهم حيث أن ميزة أخرى للتطبيق هي عداد الكلمات.
3. السرعة بالتصحيح حيث يستغرق معالجة النص باستخدام التطبيق 56 ميللي ثانية.
4. مجاني.

<u>المساوئ:</u>

1. لا يقوم بتصحيح الكلمات التي تحتوي على خطأ بأكثر من ثلاث أحرف.
2. لا يحفظ المقالات التي صححها للمستخدم من قبل.

writing9:

يساعد التطبيق الطلاب على التدرب على قسم الكتابة لامتحان IELTS وتحسين مهاراتهم الكتابية.

<u>الميزات:</u>

1. تحقق من عدد غير محدود من المقالات
2. يعطي اقتراحات للحصول على درجة أعلى
3. تصفح المقالات المصححة من قبل
4. يستخدم مدقق نحوي و إملائي دقيق ويعطي نتيجة تصحيح دقيقة
5. سريع بإعطاء النتيجة

المساوئ:

1. غير مجاني حيث تبلغ تكلفة الاشتراك 12 دولار شهرياً

Essay Grader:

الميزات:

1. تدقيق نحوي ولغوي
2. تقييم قوة الأسلوب
3. التحقق من السرقة الأدبية
4. مجاني

المساوئ:

1. يأخذ وقتاً طويلاً بالمعالجة
2. النتيجة غير دقيقة

2.2 - أبحاث مشابهة:

في الورقة [1] تم بناء عدة نماذج لشبكات عصبونية عميقة قادرة على تمثيل كل من المعلومات المحلية المتعلقة بالسياق والمستخدمة للتنبؤ بدرجات المقال منها:

نموذج (C&W) المعزز وهو عبارة عن بنية شبكة عصبونية تتعلم التمثيل الموزع لكل كلمة في مجموعة بناءً على سياقها المحلي ومن ثم تم توسيع هذا النموذج من أجل الحصول على التضمينات الخاصة بال الكلمات (SSWEs) لالتقاط معلومات تساعد أكثر من غيرها في الحصول على درجة المقال.

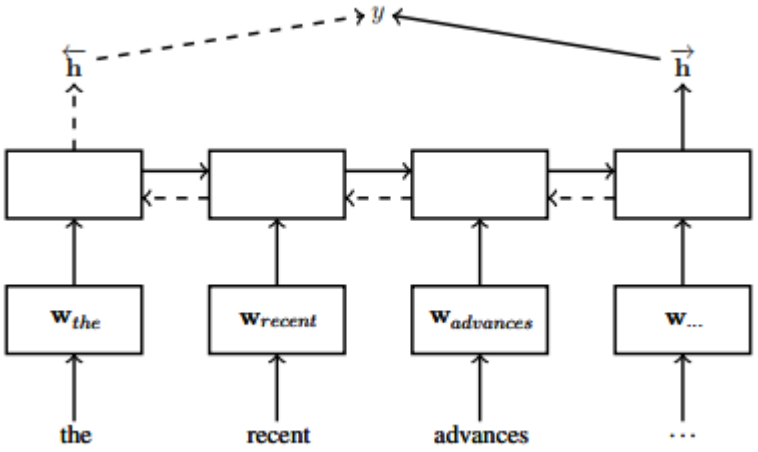
وذلك ليس لتسليط الضوء على البيئة اللغوية المحلية لكل كلمة فقط بل لمعرفة مقدار مساهمة كل كلمة بإجمالي درجة المقال.

بعد ذلك تم الاستفادة من SSWEs التي تم الحصول عليها بواسطة النموذج السابق لاشتقاق تمثيلات مستمرة لكل مقال.

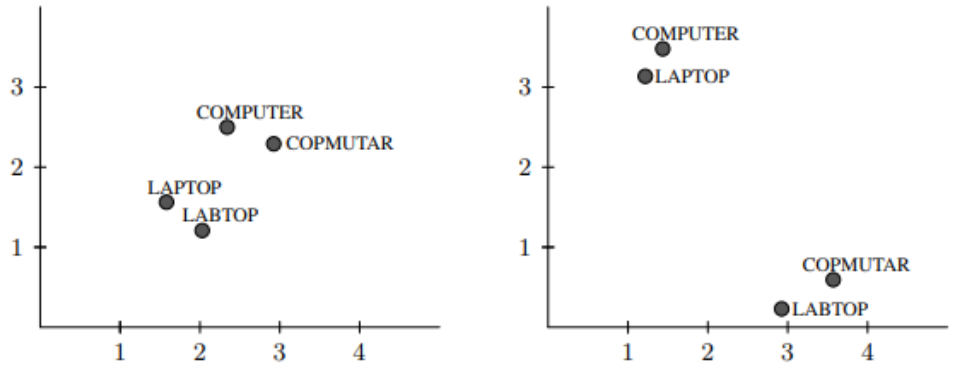
تم التعامل مع كل مقال على أنه سلسة واستخدام كل من LSTM أحادية الاتجاه وثنائية الاتجاه بشكل فعال لتضمين التسلسلات الطويلة.

LSTMs هي نوع من بنية الشبكة العصبية الإرجاعِيَّة (RNN) حيث تكون المخرجات في الزمن t معتمدَة على المدخلات في كل من الزمن t وفي الزمن t-1.

الشبكة ذات الذاكرة الطويلة قصيرة المدى ذات الطبقة الواحدة (LSTM):



متجهات الكلمة تدخل طبقة الدخل واحدة تلو الأخرى. تُستخدم الطبقة المخفية التي تم تشكيلها في الخطوة الزمنية الأخيرة للتنبؤ بدرجة المقال باستخدام الانحدار الخطي.
استخدام LSTMs ثنائي الاتجاه (الأسهم المتقطعة) يساعدنا بالحصول على تمثيلات أعمق للكلمات، يمكننا تكديس المزيد من طبقات LSTM بعد الطبقة المخفية. قد يؤدي تدريب LSTM بطريقة أحادية الاتجاه (أي من اليسار إلى اليمين) إلى استبعاد معلومات مهمة حول الجملة. على سبيل المثال، قد يختلف تفسيرنا للكلمة في نقطة ti ما عندما نعرف الكلمة عند ti + 5. الطريقة الفعالة للتغلب على هذه المشكلة هي تدريب LSTM بطريقة ثنائية الاتجاه. يتطلب ذلك القيام بتمريرة للأمام وللخلف في التسلسل (أي تغذية الكلمات من اليسار إلى اليمين ومن اليمين إلى اليسار).



(a) Standard neural embeddings

(b) Score-specific word embeddings

بالمقارنة بين تضمينات الكلمات القياسية والتضمينات المرتبطة بالمقال. فإن التضمينات العصبية القياسية ستضع الكلمات الصحيحة وغير الصحيحة بالقرب من مساحة المتجه. ونظرًا لوجود الأخطاء في المقالات ذات الدرجات المنخفضة، فإن SSWEs قادرة على التمييز بين الكلمات الصحيحة وغير الصحيحة دون خسارة في المعنى السياقي.

تم تجربة نماذج أساسية أخرى كتدريب نموذج انحدار المتّجه الداعم Support Vector Regression ، وهو أحد الأساليب الأكثر استخدامًا في تقييم النصوص. حيث تم بتحليل البيانات باستخدام محلل RASP واستخراج عدد من السمات المختلفة لتقييم جودة المقالات. وبشكل أكثر تحديدًا، تحديد أقسام الكلام (الأحادية والثنائية والثلاثية) حيث تم استبدال كل كلمة بقسم الكلام الخاص بها؛ بالإضافة إلى ذلك، تم استخراج القواعد من شجرة التركيب اللغوي واستخدامها كميزات بناءً على التحليل العلوي لكل جملة، بالإضافة إلى تقدير معدل الخطأ بناءً على القواعد الخطأ المشتقة يدويًا.

تُوزن Ngrams باستخدام (tf – idf)، في حين أن الباقي يعتمد على التكرار ويتم قياسه بحيث يكون لجميع الميزات نفس الترتيب من حيث الأهمية تقريبًا. يتم تطبيع متجهات الإدخال النهائية للوحدة لحساب الإنحيازات المتغيرة في طول النص.

يأخذ PV-DM متجهات الكلمة كدخل له التي تشكل متسلسلات ngram وتستخدم تلك لتوقع الكلمة التالية في التسلسل. ومع ذلك، فإن ميزة

PV-DM هي أنه يتم تعيين متجه فريد لكل فقرة يتم استخدامه في التنبؤ. وبالتالي، فإن هذا المتجه يعمل بمثابة ذاكرة، حيث يحتفظ بالمعلومات من جميع السياقات التي ظهرت في هذه الفقرة. يتم بعد ذلك تغذية متجهات الفقرة إلى نموذج الانحدار الخطي للحصول على درجات المقالة ويعرف هذا النموذج باسم (doc2vec).

بالإضافة إلى ذلك، تم عرض تأثير طريقتهم الخاصة بتعلم تضمينات الكلمات، عند مقارنتها بثلاثة أنواع مختلفة من embeddings الكلمات:

- تضمينات word2vec المدربة على مجموعة التدريب.
- تضمينات word2vec المتاحة للعامة المدربة مسبقًا على مجموعة أخبار Google (حوالي 100 مليار كلمة)، والتي كانت فعالة جدًا في التقاط المعلومات السياقية فقط.
- التضمينات التي يتم إنشاؤها على الفور بواسطة LSTM، من خلال إنتشار الأخطاء من طبقتها المخفية إلى مصفوفة التضمين بالتالي (لم نقدم أي عمليات تضمين للكلمات مدربة مسبقًا).

في الورقة [2]، استُخدم الانحدار الخطي للتنبؤ بدرجة المقال تلقائيًا بناءً على استخراج نوعين من السمات:

سمات بسيطة (Dense Features):

عدد المحارف والكلمات والأخطاء الإملائية وعلامات الترقيم والكلمات الصعبة وعدد ال stop words كلاً على حدى.

سمات أعمق بتحديد أقسام الكلام (part-of-speech n-grams):

وذلك بتحديد ثنائيات وثلاثيات أقسام الكلام، ويكون عدد مرات ورود كل ثنائية وثلاثية خرج هذه السمة وهي من أكثر السمات تأثيراً في عملية التقييم للمقال.

يُحدد المنظِّم L1 norm والمعرُوف أيضاً ب Lasso أكثر السِّمات فعاليةً في توقُّع الدرجات على سبيل المثال، إذا كانت سمتان مترابطتان بشكل كبير، فسيتم تعيين قيمة إحدى وزنيهما إلى الصفر، حتى لو كانت كلتا الميزتين تساعدان في عملية توقع العلامة. سيساعد ذلك في تقليل أخطاء التوقع بشكل كبير والحصول على نتائج يمكن مقارنتها بالنتائج البشرية.

في الورقة [3] تم بناء نموذج انحدار خطي لتصنيف المقالات الآلي جنبًا إلى جنب مع التركيز البسيط على التحليل الدلالي الكامن (LSA) للعثور على التماسك للمقال، سيكون الهدف من المصحح هو إعادة صياغة درجة المقال بأكبر قدر ممكن من الدقة. تقوم الآلة بتعيين الدرجات ويجب أن تكون قريبة من درجة المصحح البشري. تم استخدام مجموعة بيانات من حوالي 2000 مقالة. تم تصنيف المقالات على أساس معايير مثل عدد الكلمات، والتماسك، وعدد الكلمات الطويلة، والقواعد الصحيحة، وما إلى ذلك. يتم استخدام اختيار الميزة للوصول إلى التنبؤ بالدرجات الأكثر دقة. في النموذج يوجد أربع طبقات طبقة تمثيل المدخلات تقوم بتمثيل كل مقال للطالب على أنه ناقل، طبقة عنونة الذاكرة يتم اختيار عينة من مقال الطالب لكل درجة تم تقديرها بنفس الدرجة يمكن هنا استخدام معرفة الخبراء لاختيار العينة الأكثر تمثيلاً لكل درجة، يتم إرسال جميع العينات المختارة إلى الذاكرة كمصفوفة من المتجهات ويتم حساب وزن وأهمية كل عينة بمنتج نقطي بين وبعد، طبقة قراءة الذاكرة حيث توجد تُستخدم مصفوفة للاحتفاظ بالتمثيل المرجعي لمساحة الميزة وتُستخدم مصفوفة للتدريب لتحقيق أداء أفضل. يوضح متجه الوزن مدى تأثير العينة على النتيجة الإجمالية، طبقة الإخراج حيث عبارة عن مصفوفة، وهو عدد الدرجات المتوقعة وهي قيمة التحيز. كمية العقد الناتجة تساوي طول نطاق النتيجة. وبالتالي، من خلال حساب أعلى درجة محتملة من توزيع الدرجات المحتملة هو كيفية وصولنا إلى قيمة التقدير النهائية لعرضها لدرجة العينة المحددة.

تم اختبار النموذج على مجموعة بيانات ASAP وحقق درجة كبيرة في 6 من أصل 8 مجموعات بيانات.

وبالتالي يمكن استخدام هذا النموذج عندما يكون الوقت جوهريًا أو لمنح الطلاب معاينة للدرجة التي قد يحققونها قبل تسليم أوراقهم. لتحسين النموذج بشكل أكبر، يمكننا عمل تمثيل متجه للتعلم للمهمة. وبالتالي يمكننا في النهاية ربط تمثيل المتجه بالنتيجة.

ومع ذلك، إن النموذج قد تم اختباره على مجموعة بيانات واحدة فقط، وقد يكون له تحيزات قد تؤثر على المخرجات.

في الورقة [4] نقترح تقنية جديدة للتكييف مع المجال تعتمد على انحدار التلال الخطي Bayesian.

تم استخدم نظام تسجيل مقال مفتوح المصدر، EASE (محرك درجات الذكاء الاصطناعي المحسّن)، لاستخراج السِّمات. يتم تعريف n-grams المفيدة على أنها n-grams التي تنقسم إلى مقالات ذات درجات جيدة وتقول الدرجات السيئة، ويتم تحديدها باستخدام اختبار Fisher. المقالات ذات الدرجات الجيدة هي مقالات ذات درجة أكبر من متوسط الدرجات أو تساويها ، والباقي تعتبر مقالات ذات درجات سيئة. ثم يتم اختيار أعلى 201 ن-جرام مع أعلى قيم فيشر كميزات الحقيبة. يستخدم EASE ال NLTK لوضع علامات على POS والاشتقاق، و aspell للتدقيق الإملائي، و WordNet للحصول على المرادفات. يتم إنشاء POS tagging الصحيحة باستخدام نص صحيح نحويًا (يتم توفيره بواسطة EASE). يستخدم برنامج EASE ال scikit-Learn لاستخراج ميزات unigram و bigram. بالنسبة للانحدار الخطي، يتم إلحاق ميزة ثابتة للقيمة الأولى للانحياز.

بالنسبة للنتائج: نحدد الحدود العليا الإرشادية على درجات QWK باستخدام انحدار التلال الخطي(Bayesian BLRR). تحقيقا لهذه الغاية ، نقوم بإجراء التحقق من صحة 5 أضعاف من خلال التدريب والاختبار داخل كل مجال.

يتم ذلك باستخدام انحدار آلة متجه الدعم الخطي (SVM) لتأكيد أن BLRR هي طريقة تنافسية لهذه المهمة. بالإضافة إلى ذلك ، نظرًا لأن بيانات ASAP تحتوي على اثنين على الأقل من الشروح البشرية لكل مقال ، فإننا نحسب أيضًا درجة الاتفاق البشري. النتائج. درجات BLRR قريبة من درجات الاتفاق البشري للموجه 1 والمطالبات من 5 إلى 8 ، ولكنها تقصر بنسبة 10٪ إلى 20٪ للمطالبات من 2 إلى 4. ونرى أيضًا أن يمكن مقارنة BLRR بانحدار SVM الخطي، مما يعطي نفس الأداء تقريبًا للمطالبات من 4 إلى 7 ؛ أداء ضعيف قليلاً للمطالبات من 1 إلى 3 ؛ وأداء أفضل بكثير للموجه 8.

في الورقة [5] تم إنشاء نظامًا آليًا لتسجيل درجات المقالات، كان التركيز في تصنيف المقالة على أسلوب المقال، الذي قاموا بتوسيعه بإضافة فئة الادراك. قاموا بتقييم الانحدار الخطي، وشجرة الانحدار، والتحليل التمييزي الخطي، وآلات المتجهات الداعمة على الميزات وتم اكتشاف أن SVM حقق أفضل النتائج بمتوسط $\kappa = 0.78$.

في الورقة [6] تم النظر إلى خمس فئات من الميزات التي تم النظر فيها وإنشاءها لهذا المشروع من أجل الأسلوب وهي طبيعتها البصرية، إدراج الضمائر، وكلماتها الجميلة حيث فقط الكلمات التي يزيد طولها عن 5 أحرف تعتبر جميلة، فعاليتها العاطفية، وإدراكها.

بالنسبة لخوارزميات التعلم تم تجريب كل من :

الانحدار الخطي، SVM، تحليل التمييز الخطي متعدد الطبقات، أشجار الانحدار.

التي تليها في الثانية. Naive Bayes مع 1 أفضل أداء على مجموعة المقالات SVMكان أداء بالنسبة لهاتين الخوارزميتين، يبدو أن خطأ التدريب يتناقص مع زيادة حجم مجموعة التدريب وهو أمر مرغوب فيه.

في [7] قام BERT و RoBERTa بوضع أداء جديد متطور في مهام الانحدار بين الجمل مثل التشابه النصي الدلالي (STS). ومع ذلك، فإنه يتطلب إدخال كلتا الجملتين في الشبكة، مما يسبب نفقات حسابية هائلة يتطلب العثور على الزوج الأكثر تشابها في مجموعة من 10000 جملة حوالي 50 مليون عملية حسابية استدلالية (~ 65 ساعة) باستخدام BERT .

إن بناء BERT يجعله غير مناسب للبحث عن التشابه الدلالي وكذلك للمهام غير الخاضعة للإشراف مثل التجميع.

لكن هنا نقوم بتقييم Sentence-BERT (SBERT) ، وهو تعديل لشبكة BERT المدربة مسبقا والتي تستخدم هياكل الشبكة السيامية Siamese والثلاثية لاشتقاق تضمينات الجملة ذات المعنى الدلالي والتي يمكن مقارنتها باستخدام تشابه جيب التمام. هذا يقلل من الجهد المبذول للعثور على الزوج الأكثر تشابها من 65 ساعة مع BERT / RoBERTa إلى حوالي 5 ثوان مع SBERT، مع الحفاظ على الدقة من BERT.

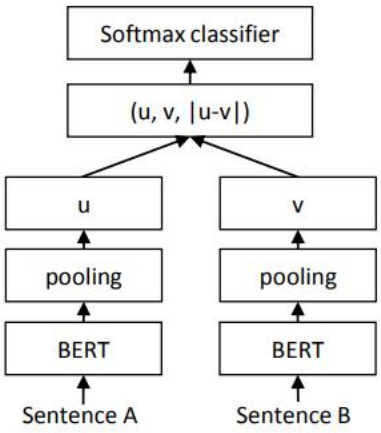


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

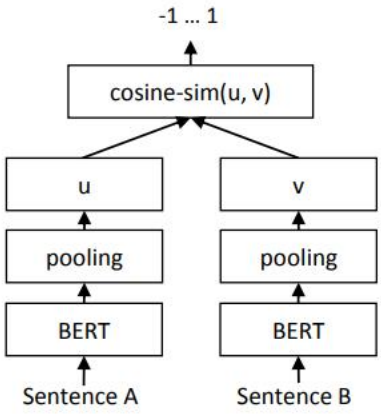


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

في الورقة [8] اقتُرحت نماذج تعتمد على الشبكات العصبية الإرجاعِيَّة لمعالجة مهمة التقييم الآلي للمقال. لا تعتمد الطريقة على أي هندسة للميزات بل تتعلم تلقائيًا التمثيلات المطلوبة للمهمة. لقد تم اكتشاف مجموعة متنوعة من هياكل نماذج الشبكة العصبية لتقييم المقالات آلياً وتم تحقيق تحسينات كبيرة بالنسبة لنظام أساسه قوي. يتفوق أفضل نظام على النظام الإبتدائي الأساسي بنسبة 5.6٪ فيما يتعلق بالموزون التربيعي Kappa (QWK) . علاوة على ذلك ، تم إجراء تحليل للشبكة للحصول على نظرة ثاقبة لنموذج الشبكة العصبية الإرجاعيَة واتضح أن الطريقة تستخدم بشكل فعال محتوى المقالة لاستخراج المعلومات المطلوبة لتقييم المقالات.

الفصل الثالث
الدراسة التحليلية

الدراسة التحليلية

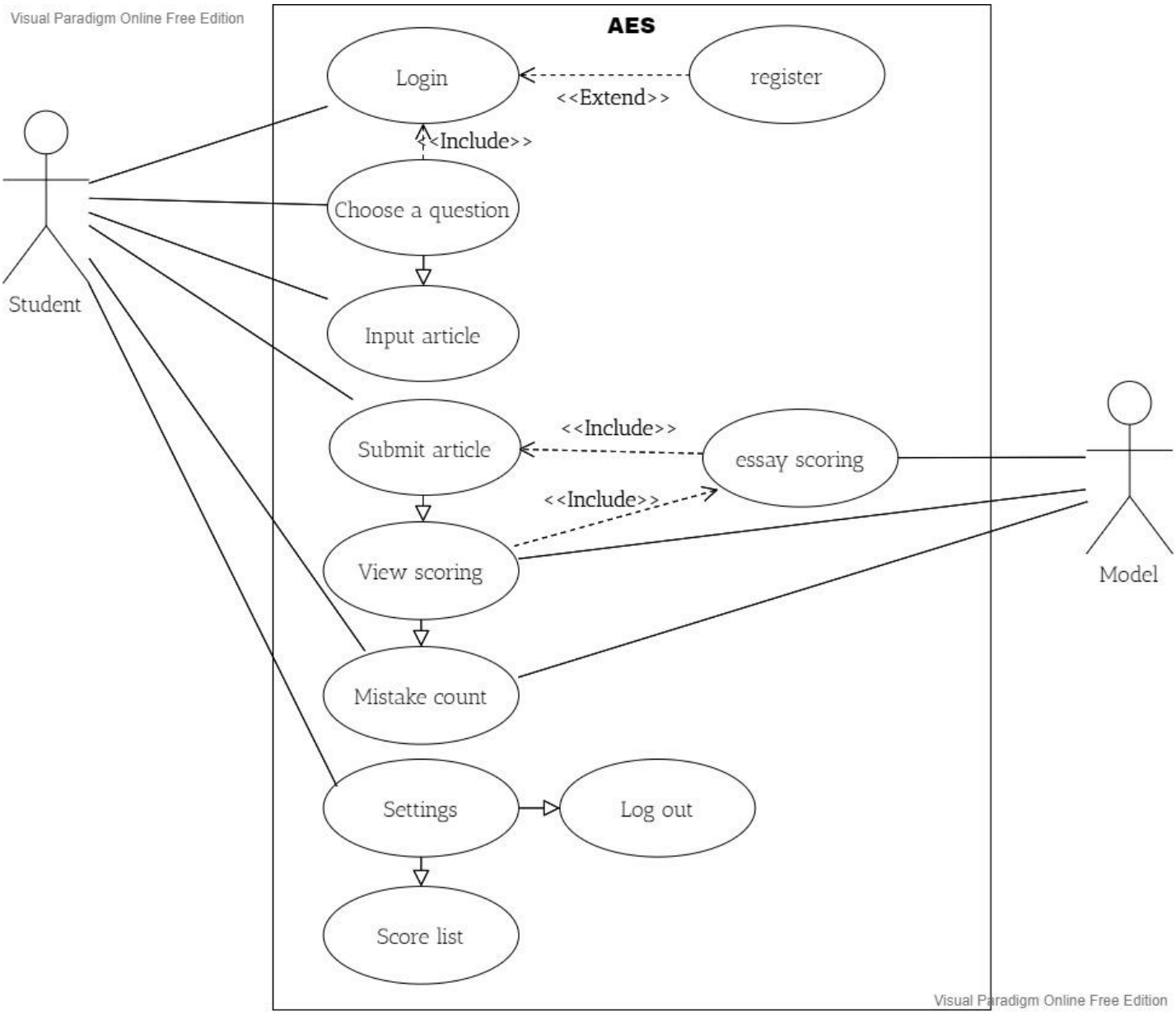
3.1 المتطلبات الوظيفية:

- تقديم موضوع معين لمناقشته ضمن مقال يكتبه المستخدم.
- تقييم المقال و إعطاء علامة تقريبية.
- حفظ مقالات المستخدم السابقة مع نتيجة التقييم التي حصل عليها.

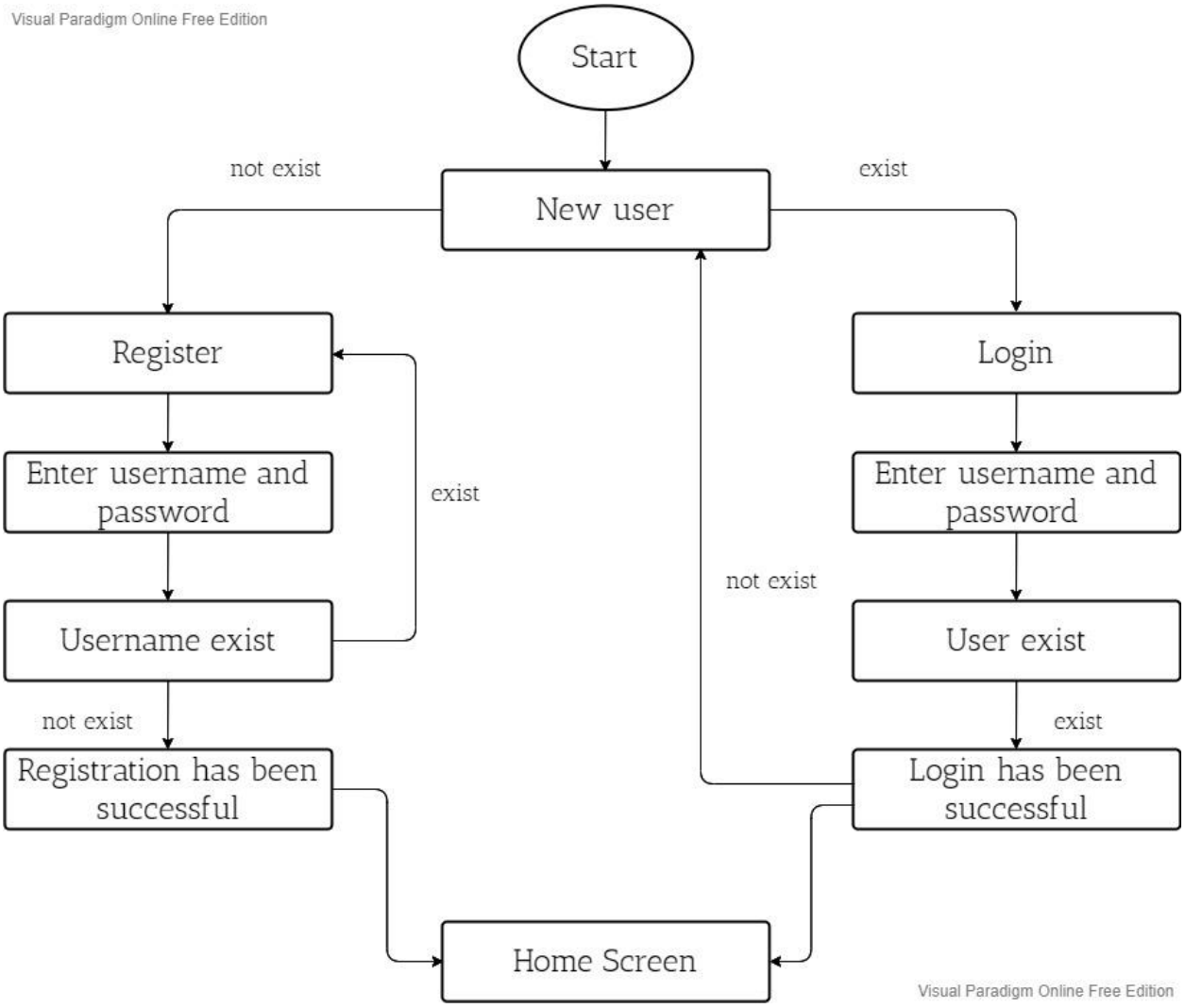
3.2 المتطلبات غير الوظيفية:

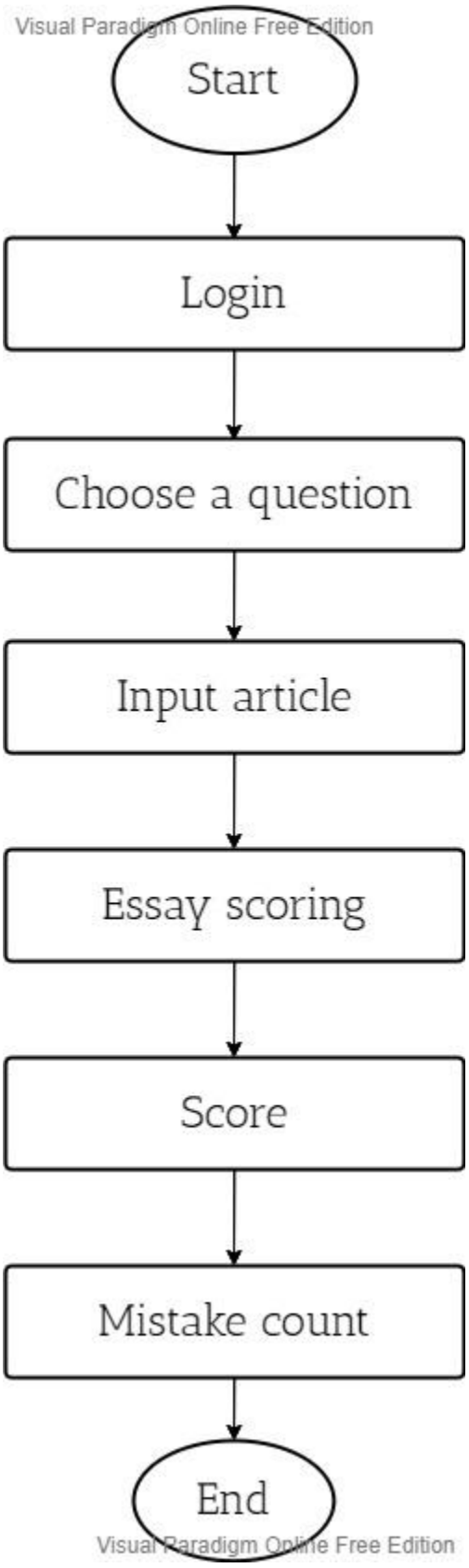
- سهولة الاستخدام
- سهولة التعديل والصيانة
- واجهات سهلة الاستخدام ومريحة

3.3 حالات الاستخدام:



3.4 Flow charts:




Visual Paradigm Online Free Edition
Start
Login
Choose a question
Input article
Essay scoring
Score
Mistake count
End
Visual Paradigm Online Free Edition

الفصل الرابع:
التجارب والاختبار

التجارب و الاختبار

4.1 - نموذج تقييم المقال:

LSTM+Word2Vec:

قمنا بمعالجة المقالات حيث تم إزالة الstop words وعلامات الترقيم ثم قمنا بإنشاء قائمة بالكلمات من كل جملة ومن كل مقال. يتم إنشاء الميزات من خلال تمرير المقالات لنموذج Word2Vec يعمل النموذج Word2Vec كطبقة Embedding في شبكة عصبية.

يتم تمرير الميزات من هذا الطراز من خلال طبقات LSTM الخاصة بنا. نحن ننفذ طبقتين LSTMتقبل الطبقة الأولى جميع الميزات من (Word2Vec)كمدخل وتمرر 300 ميزة كخرج إلى الطبقة الثانية من LSTM تقبل الطبقة الثانية 300 ميزة كمدخل و 64 ميزة كخرج. بعد ذلك نضيف طبقة Dropout بقيمة 0.5. أخيرًا طبقة Dense مع خرج 1 الذي يمثل درجة المقال.

تم تدريب النموذج على 150 حقبة بحجم دفعة 64 مع استخدام الearly stopping

النتائج:

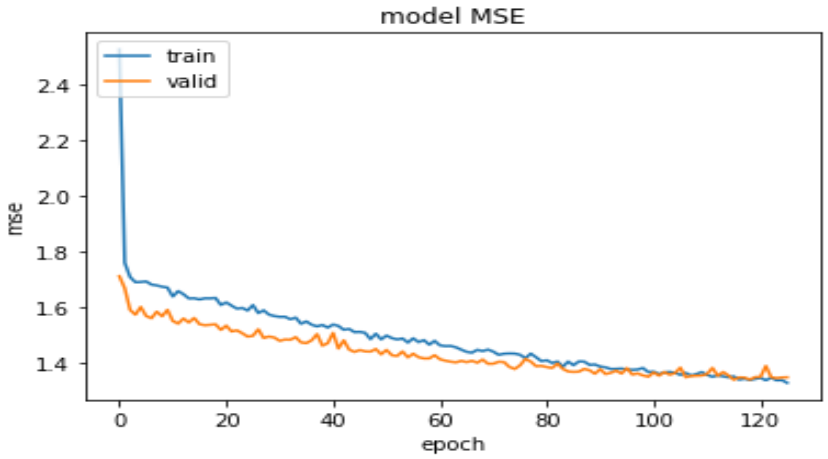
MAE: 1.304149852008054

MSE: 2.91571015322583

RMSE: 1.707545066235685

Pearson correlation: 0.682

kappa:0.61



Machine learning model for each topic:

لجعل بياناتنا جاهزة لتطبيق الخوارزميات، نحتاج إلى خطوة أخرى. لا يمكن تطبيق خوارزميات التعلم الآلي على الجمل أو الكلمات، ولا يمكن استخدامها إلا على البيانات الرقمية. تحتوي مجموعة البيانات الخاصة بنا على مجال يحتوي على مقالات تحتاج إلى تحويلها إلى شكل رقمي أولاً من أجل تدريبها. للقيام بذلك، قمنا باستخراج السمات التي وجدناها مناسبة مثل:

عدد الأخطاء الإملائية والقواعدية، عدد الكلمات الغير شائعة، عدد الجذور، عدد الكلمات، عدد الجمل، عدد علامات الترقيم، عدد الأفعال، عدد الأسماء، عدد الصفات، عدد الكلمات الصعبة، ميزات قابلية القراءة للنص.

في البداية قمنا بتطبيق خوارزميات التعلم الآلي مثل الانحدار الخطي و XGBRegressor و Random Forest و ElasticNetو GradientBoostingRegressorعلى مجموعة البيانات مع عمل scale للداتا.

حيث حقق الElasticNet أفضل النتائج لأن الخوارزميات الباقية مالت للoverfit

قمنا بتمرير كل قسم من الداتا على حدا لgrid search ليتم اختيار أفضل البرامترات بالنسبة لكل قسم.

وبالتالي كانت النتائج:

	MAE	MSE	R2	RMSE	Kappa
Topic 1	0.64	0.69	0.70	0.83	0.81
Topic 2	0.75	0.89	0.56	1.03	0.69
Topic 3	1.55	4.14	0.48	2.0	0.64
Topic 4	1.67	4.57	0.51	2.13	0.67
Topic 5	1.02	1.77	0.71	1.33	0.79
Topic 6	1.26	2.44	0.59	1.56	0.71
Topic 7	0.77	0.94	0.58	0.97	0.71
Topic 8	0.44	0.30	0.46	0.55	0.55

Machine learning model for all topics:

مرحلة التدريب:

✓ بداية تم تدريب السمات ال 46 (الموجودين بقسم التحقيق)

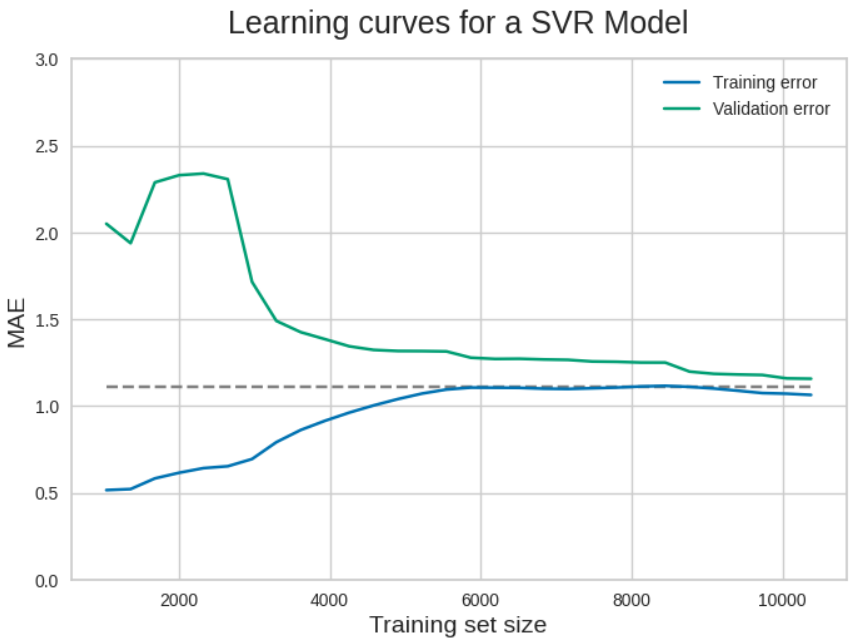
النماذج المستخدمة:

Support Vector Regressor (SVR) -1:

أفضل المتغيرات بعد تطبيق الgridSearsh:

```
SVR(kernel='rbf',C= 100, epsilon =0.7, gamma =0.004)
```

النتائج:



4.2 - نموذج ارتباط الموضوع بالمقال قمنا بثلاث خطوات:

1. تلخيص المقال الى اهم الافكار البارزة فيه
2. قياس مدى ارتباط الموضوع مع المقال الملخص
3. حساب المتوسط الحسابي لنسب التشابه الناتجة بين الموضوع والجمل الملخصة

1- تلخيص المقال الى أهم الافكار البارزة فيه:

نحن نحتاج لقياس مدى ترابط الموضوع بالمقال الى معرفة الفكرة الاساسية من المقال بالنسبة للموضوع ولكن يملك المقال كمية هائلة من البيانات النصية، وهناك حاجة كبيرة لتقليل الكثير من هذه البيانات النصية إلى ملخصات أقصر ومركزة تلتقط التفاصيل البارزة، حتى نتمكن من قراءتها والتنقل خلالها بسهولة لذلك قمنا بعملية التلخيص التلقائي للنص.

التلخيص التلقائي للنص هو مهمة تكثيف جزء من النص إلى نسخة أقصر، مع الحفاظ على النقاط المهمة ومعنى المحتوى وذلك من خلال :

- إنشاء مصفوفة التشابه:

سيتم إنشاء مصفوفة التشابه عن طريق حساب درجة التشابه للجملة فيما يتعلق بجميع الجمل الأخرى الموجودة في نص الإدخال. ضع في اعتبارك أن نص الإدخال يتكون من n جملة، ثم سيتم تمييز أبعاد مصفوفة التشابه على أنها S1، S2، ...، Sn. الآن ، سنأخذ جملتين Si و Sj ، لحساب درجة التشابه للجملة Si مع الجملة Sj وبالمثل ، سنحسب درجة التشابه لجميع الجمل الأخرى ونقوم بتحديثها في مصفوفة التشابه. في حالة وجود الجمل نفسها، ستكون درجة التشابه صفرًا.

- ترتيب الجمل:

بعد إنشاء مصفوفة التشابه، نحتاج إلى ترتيب الجمل بناءً على درجة التشابه الفعلية. درجة التشابه الفعلية هي مجموع كل درجات التشابه للجملة فيما يتعلق بجميع الجمل الأخرى الموجودة في نص الإدخال. الآن، تم ترتيب الجمل بترتيب تنازلي بناءً على درجة التشابه الفعلية.

- استخراج الجملة:

لإنشاء الملخص لنص الإدخال المحدد ، سنقوم باستخراج الجمل العلوية بناءً على ترتيبها من الخطوة السابقة. سيتم بعد ذلك إعادة ترتيب الجمل المستخرجة بترتيب زمني. بعد ذلك، سيتم اعتباره ملخصًا مقتطفًا لنص الإدخال المحدد ، وسيتم تمريره إلى المرحلة التالية كنص إدخال.

2- قياس مدى ارتباط الموضوع مع المقال الملخص:

سوف نقوم بحساب درجة التشابه للموضوع بالنسبة لجميع الجمل الأخرى الموجودة في نص الإدخال والتي هي (المقال الملخص) ونقوم بجمع النسب

3- حساب المتوسط الحسابي لنسب التشابه الناتجة بين الموضوع والجمل الملخصة :

حيث نقوم بأخذ الناتج عن عملية قياس مدى ترابط الموضوع بالمقال الملخص وتقسيمها على عدد الجمل فنحصل على نسبة التشابه التي نريدها

تساعد عملية مقارنة الموضوع بأكثر من جملة من زيادة تأكيد ان هذا الموضوع يرتبط بهذا المقال حيث ان وجود جملة غير مرتبطة بالموضوع تقلل من قيمة الناتج عند تقسيمها على عدد الجمل وبذلك نكون قد حصلنا على قيمة الارتباط بأدق ما يمكن.

لقياس عملية التشابه قمنا بتجريب ثلاث توابع :

1. Cosin_distance[6]: يستخدم تشابه جيب التمام لتحديد التشابه بين المستندات أو المتجهات. رياضيا ، يقيس جيب التمام للزاوية بين متجهين في فضاء متعدد الأبعاد.

2. Sentence Similarity: هو مهمة تحديد مدى تشابه النصين. تقوم نماذج تشابه الجمل بتحويل نصوص الإدخال إلى متجهات (embeddings) تلتقط المعلومات الدلالية وتحسب مدى قربها (تشابهها) بينها.

3. Sentence Transformers[7]: تضمين الجملة والفقرة متعدد اللغات باستخدام $BERT\&Co$ يوفر هذا الإطار طريقة سهلة لحساب تمثيلات المتجهات الكثيفة للجمل والفقرات والصور .تعتمد النماذج على شبكات المحولات مثل BERT / RoBERTa / XLM-RoBERTa وما إلى ذلك وتحقيق أداء متطور في مختلف المهام. يتم تضمين النص في الفضاء المتجه بحيث يكون النص المماثل قريبا ويمكن العثور عليه بكفاءة باستخدام تشابه جيب التمام.

بالنسبة لتوابع التشابه لم نستطع تعميم تابع على كل الحالات حيث في حالات استطاع تابع معالجتها لم يستطع تابع اخر فعل ذلك حسب نوع الموضوع و المقال وعدد كلماته وصعوبة هذه الكلمات

ولكن تم الاعتماد على Sentence Transformers لان نتائجها كانت الافضل

عندما تكون المقال للموضوع تكون المتوسط يزيد عن 3:

sum value 0.6805827945978276

avg value 3.402913972989138

عندما تكون المقال لغير الموضوع تكون النتائج اقل من 3:

sum value 0.5044579675312015

avg value 2.5222898376560074

الفصل الخامس:
التحقيق

التحقيق

ذكرنا في أقسامٍ سابقةٍ معماريّة النظام، حيث قمنا بسرد الأجزاء الأساسيّة التي يتكوّن منها وكيفيّة تفاعلها مع بعضها البعض، بشكلٍ يؤمّن المتطلبات الوظيفيّة وغير الوظيفيّة المتفق عليها، وأمّا في هذا القسم، فسنقوم بذكر الطريقة التي قمنا عبرها بتحقيق كل مكوّن من هذه المكونات عبر سرد التقنيّات التي استخدمناها مع شرحٍ كافٍ للطريقة التي استُخدِم بها.

- بنية النظام:

5.1 - الموارد المستخدمة:

-الموارد العتاديّة:

حواسيب محمولة ومجهزة بنظام تشغيل Microsoft Windows 10

-الموارد البرمجية:

البيئة المستخدمة:

Android application with flutter

- اللغة المستخدمة:

Python و flask لبناء السيرفر.

5.2 - بيانات التدريب:

أقامت مؤسسة Hewlett عام 2012 مسابقة على منصة Kaggle ، تحت اسم جائزة التقييم الآلي، لإظهار قوة التصحيح الآلي، واجتذاب مطوري الأنظمة الجديدة، قام المشاركون بوضع خوارزميات تصحيح آلية تتنبأ بالعلامة التي سيعطيها مصحح بشري لإجابة تحريرية، وشملت المقالات على 12.976 إجابة موزعة في ثمانِ فئات ذات مواضيع مختلفة تتراوح من 150 إلى 550 كلمة لكل منها، ولكلٍّ منها سِمَات وطريقة تصحيح ونطاق درجات مختلف عن الأخر.

كتب المقالات طلاب من المرحلة الإعدادية تتراوح أعمارهم بين الصف السابع والعاشر، كما كانت المقالات مصحَّحة من قبل مصحِّحين بشريِّين.

تم تقسيم المقالات بشكل أساسي إلى نوعين:

الردود المعتمدة على نص معطى، والردود المقنعة / السردية / التفسيرية.

الردود المعتمدة على نص معطى: عبارة عن نص يجب على الطلاب قراءته أولاً قبل الإجابة.

الردود المقنعة / السردية / التفسيرية: تطلب من الطلاب تقديم قصص أو حكايات أو حجج رسمية لإقناع القارئ بالاتفاق مع رأي الطالب في موضوع معين.

تم إصدار مجموعة الاختبار بدون العلامة الفعلية، وبالتالي فإننا مقيدون بتقسيم مجموعة التدريب المحددة لإنشاء مجموعة اختبار جديدة.

Table 1: Dataset statistics

Essay Set	Essay Type	Domain	Score Range	Average Length	train	dev.	test	total
1	Persuasive/Narrative/Expository	-	2-12	350 words	1,284	321	178	1,783
2	Persuasive/Narrative/Expository	Writing Applications	1-6	350 words	1,296	324	180	1,800
		Language Conventions	1-4					
3	Source Dependent Responses	-	0-3	150 words	1,244	310	172	1,726
4	Source Dependent Responses	-	0-4	150 words	1,276	319	177	1,772
5	Source Dependent Responses	-	0-4	150 words	1,300	325	180	1,805
6	Source Dependent Responses	-	0-4	150 words	1,296	324	180	1,800
7	Persuasive/Narrative/Expository	-	0-30	250 words	1,131	282	156	1,569
8	Persuasive/Narrative/Expository	-	0-60	650 words	521	130	72	723

5.3 - معايير الأداء:

تم تقييم النتائج بناءً عدة معايير وهي:

Quadratic Weighted Kappa

وهو معيار شائع في مسابقات Kaggle يعبر عن مدى توافق درجات المقال التي سيتنبأ بها النظام مقارنةً بدرجات المقيم البشري.

بالنسبة للدرجات المحتملة لكل مقال N، يتم إنشاء المصفوفة بأبعاد $N \times N$ حيث تمثل عدد المقالات التي تلقت الدرجة i من المقيم البشري والدرجة j من النظام.

يتم حساب مصفوفة الأوزان $W_{i,j}$ ذات الأبعاد $N \times N$ بأخذ الفرق بين قيم الدرجات الفعلية والمتوقعة:

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

يتم إنشاء المصفوفة E بإفتراض عدم وجود ارتباط بين القيم، ويتم حسابها على أنها الجداء الخارجي بين المتجه التكراري الفعلي للنتائج والمتجه التكراري المتوقع.

يتم تسوية المصفوفات E و O بحيث يكون لها نفس المجموع.

وأخيراً لحساب QWK :

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j}\, O_{i,j}}{\sum_{i,j} W_{i,j}\, E_{i,j}}$$

نلاحظ أن QWK هي نسبة يمكن أن تأخذ قيمة بين -1 و 1.

تشير درجة QWK السلبية إلى أن النموذج "أسوأ من التوقع العشوائي" مما يعني أنه لايوجد اتفاق فعَّال بين المقيّمين حيث يعطي النموذج العشوائي درجة قريبة من 0.

بينما ستؤدي التنبؤات المثالية في حال إتفاق درجات المقال للمقيمين كلّياً إلى الحصول على درجة 1.

معيارRoot Mean Square Error (RMSE) :

$$RMSE\ (X,h) = \sqrt{\frac{1}{m}\sum_{i=0}^{m}\big(\ h(x^i) - (y^i)\big)^2}$$

هو الجذر التربيعي لمتوسط الخطأ التربيعي بين القيم المتوقعة والفعلية.

عادةً ما يُستخدم هذا المعيار في تقييم مسائل الانحدار (Regression). حيث يعطي فكرة عن مقدار الخطأ في التوقع، مع إعطاء أوزان عالية للأخطاء الكبيرة.

إنه امتداد لمعيار متوسط الأخطاء التربيعي (MSE) الذي يُستخدم بشكل أساسي عندما تحتوي التنبؤات على انحرافات كبيرة،نحن لا نريد معاقبة الانحرافات في التنبؤ كما هو الحال مع MSE.

معيار Mean Absolute Error (MAE) :

$$MAE\ (X, h) = \frac{1}{m}\sum_{i=0}^{m} \left| h(x^i) - (y^i) \right|$$

هو متوسط الخطأ المطلق (غير السالب) بين القيم الفعلية والمتوقعة.

بالتالي المتوسط المجمع لهذه الأخطاء، يساعدنا على فهم أداء النموذج على مجموعة البيانات بأكملها

هو معيار شائع الاستخدام في مسائل التنبؤ حيث يتم تفسير قيمة الخطأ بسهولة. وذلك لأن هذه القيمة التي يعطيها لها نفس scale الدرجات الذي نتوقعها.

الفرق بين RMSE و MAE:

في حين أن كلاهما لهما نفس الهدف وهو قياس خطأ نموذج الانحدار، إلا أن هناك بعض الاختلافات الرئيسية:

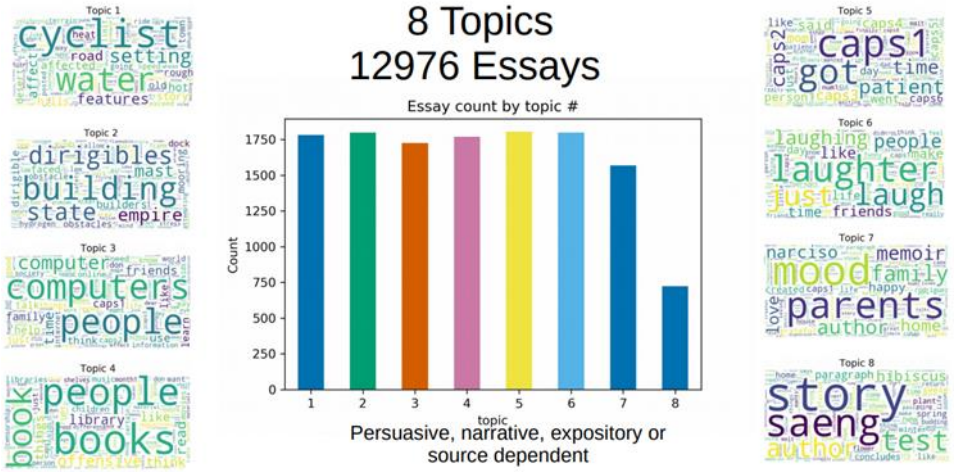
RMSE أكثر حساسية للقيم المتطرفة

تُعاقب RMSE الأخطاء الكبيرة أكثر من MAE نظرًا لحقيقة أن الأخطاء يتم تربيعها في البداية.

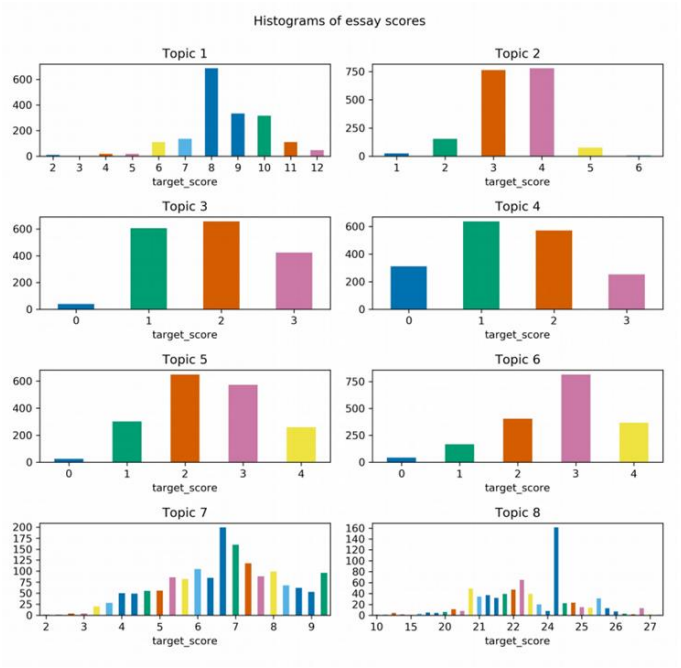
قيم MAE أكثر قابلية للتفسير لأنَّها ببساطة تعبر عن متوسط الخطأ المطلق.

تحليل واستكشاف البيانات:Exploratory Data Analysis and Topic Modeling

- تعيين موضوع يعبر عن كل صنف من الأصناف الثمانية وذلك عن طريق ال WordCloud .
- استكشاف عدد المقالات في كل موضوع من المواضيع الثمانية.



- استكشاف نطاق درجات كل موضوع على حدا.

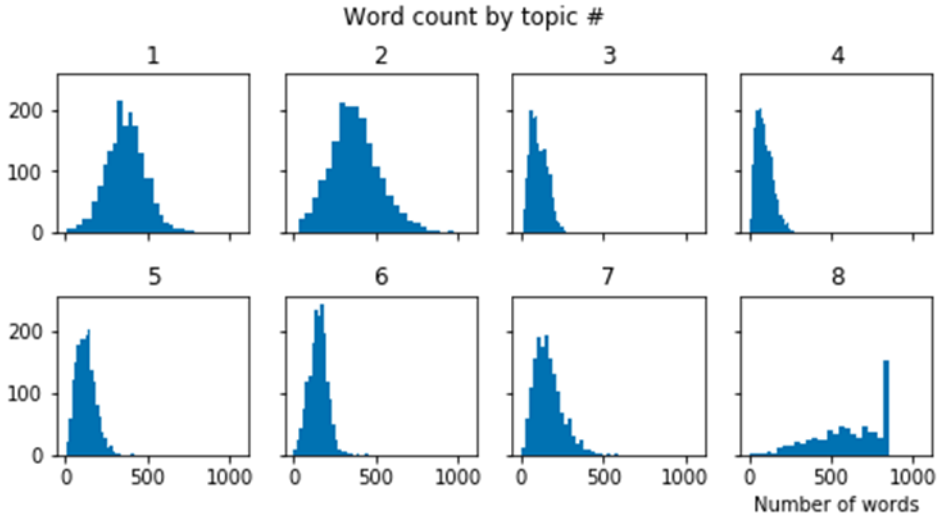


مرحلة معالجة النص الطبيعية:Natural Language Processing

- قبل البدئ بأي عملية تم التأكد من وجود أي قيم فارغةNULL .
- تحديد عدد أخطاء كل مقال وكذلك تحديد أقسام الكلام POS والتي ستساعدنا لاحقاً في استخراج سمات تفيد في توقع درجات المقال.

- كما يعتبر التعرف على أسماء العلم (NER) أحد أكثر مهام المعالجة المسبقة للبيانات حيث تتضمن تحديد المعلومات الأساسية في النص، مثل: شخص ومنظمة والمكان / الموقع والتاريخ / الوقت. التعبير القياس العددي (المال، النسبة المئوية، الوزن، إلخ) وكذلك عنوان البريد الالكتروني.

- معرفة توزع عدد كلمات كل موضوع على حدا.



- اكتشاف علاقة بعض من السمات المستخرجة في توقع درجة المقال:

عدد الكلمات/ عدد الصفات/ نسبه تشابه النص (بناءً على مقال مرجعي حاصل على درجة عالية لكل موضوع)

يتألف النظام من نموذجين أساسيين:

1. نموذج مبني باستخدام سمات مختارة يدوياً لقياس جودة الكتابة.
2. نموذج لقياس مدى ترابط المقال بالسؤال.

النموذج الأول :لقياس جودة الكتابة

سنستخدم في هذا النموذج بيانات المسابقة، تتراوح المقالات المختارة بمتوسط طول من 150 إلى 550 كلمة لكل إجابة.

السمات المستعملة:

- ✓ عدد الجمل.
- ✓ عدد علامات الترقيم: وهي عدد إشارات التعجب والاستفهام والفواصل والإقتباس كل منها على حدى.
- ✓ عدد الكلمات.
- ✓ الطول الوسطي للكلمات: ويتم حسابها بتقسيم مجموع أطوال الكلمات على عدد الكلمات.
- ✓ الطول الوسطي للجمل: ويتم حسابها بتقسيم مجموع أطوال الجمل على عدد الجمل.
- ✓ كثافة الجمل: ويتم حسابها كنسبة مئوية بتقسيم عدد الجمل على عدد الكلمات.
- ✓ عدد الأحرف.
- ✓ عدد الأفعال والأسماء والصفات والظروف والضمائر وأحرف العطف والأرقام محدِّدات الجملة الاسمية كل منها على حدا، وذلك بالاستفادة من تحديد أقسام الكلام حيث تبدأ علامة POS الخاصة بالأفعال دوماً بVB ، والصفات تبدأ بJJ ، وتبدأ الأسماء بNN ، والظروف بRB .
- ✓ عدد الأخطاء الإملائية والقواعدية: سنقوم بالتأكد من الكتابة الصحيحة لكل كلمة عن طريق language-tool-python .

- ✓ عدد الكلمات الصعبة: وهي الكلمات ال 5000 التي تعتبر صعبة ضمن امتحان SAT المعياري والموجودة ضمن الرابط التالي:

http://freevocabulary.com/

- ✓ الكلمات الشائعة: يدل استعمال الكلمات الشائعة على ثقافة أوسع وتمكن من اللغة لذلك قمنا بتضمين تواتر الكلمات الإنكليزية في كتب غوغل، باعتبار هذه البيانات شاملة للغة الإنكليزية، والتي يمكن تحميلها من الرابط التالي:

norvig.com/google-books-common-words.txt

ثم نقوم بحساب التواتر الوسطي، ونسبة هذه الكلمة في اللغة الإنكليزية، ومن ثم تقسيم الكلمات إلى شائعة وغير شائعة وذلك باعتبار كل الكلمات التي ترد بعدد مرات أقل من الوسطي كلمات غير شائعة. وعند جمع السمات، نقوم بتحديد كلمات النص الغير الشائعة وتحصيل نسبة كل منها بالجمع وهي الصورة النهائية لهذه السمة.

- ✓ ثنائيات وثلاثيات أقسام الكلام: وهي من أكثر السمات تأثيراً في عملية التقييم للمقال.
- ✓ عدد الstop words .
- ✓ عدد أسماء العلم (NER) كل منها على حدا.
- ✓ سمات سهولة القراءة حيث تشير إلى سهولة قراءة النص من أجل تحديد إمكانية قراءة النص بدقة، يجب النظر في العديد من العوامل وحسابها. يتم صياغة خوارزميات حاسوبية دقيقة لتحليل مكونات متعددة للنص بما في ذلك عدد الكلمات وطول الجملة والحروف لكل كلمة وما إلى ذلك. تم سرد العديد من طرق تسجيل سهولة القراءة الأكثر استخدامًا وشرحها أدناه حيث تحلل كل طريقة تسجيل عوامل مختلفة، مما يعني أن النتائج متنوعة أيضًا.
    - Automated Readability Index (ARI): مصممة لقياس إمكانية فهم النص حيث الناتج هو تمثيل تقريبي لمستوى الدرجة الأمريكية اللازم لفهم النص.

$$ARI = 4.71 * (characters/words) + 0.5 * (words/sentence) -21.43$$

- Flesch Reading Ease: تم تصميمه للإشارة إلى مدى صعوبة فهم مقطع القراءة حيث تشير الدرجات الأعلى إلى المواد التي يسهل قراءتها، تشير الأرقام المنخفضة إلى مقاطع يصعب قراءتها.

FRE = 206.835 − 1.015*(total words/total sentences) − 84.6 * (total syllables/ total words)

- Coleman-Liau Index: مصممة لقياس إمكانية فهم النص. الناتج هو المستوى التقريبي للدرجة الأمريكية الذي يعتقد أنه ضروري لفهم النص.

CLI = (5.89 * (characters/ words)) −(30 *(sentences/words)) − 15.8

- Gunning Fog Index: مصممة لقياس سهولة قراءة عينة من الكتابة الإنجليزية. الفهرس الناتج هو مؤشر على عدد سنوات التعليم الرسمي (الدرجة الأمريكية) التي يحتاجها الشخص من أجل فهم النص بسهولة في القراءة الأولى.

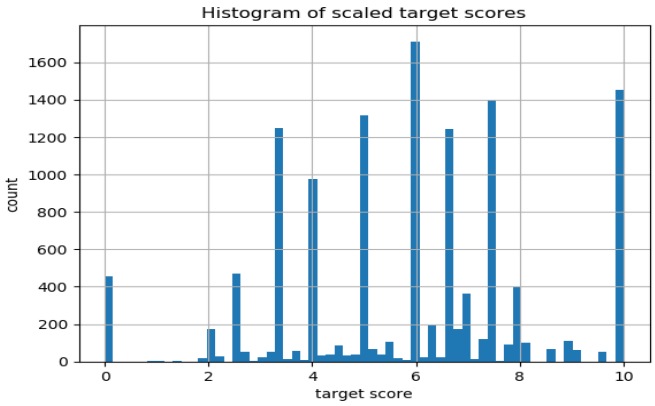
GFI = 0.4 * ((words/ sentence) +100 * (complex words/ words))

مرحلة توحيد الداتا:

بدلاً من تناسب كل موضوع على حدا نقوم بالجمع بين جميع المواضيع في نموذج واحد.

ونظرًا لأن كل موضوع يحتوي على نطاق مختلف من الدرجات، فيجب أن يتم توحيد درجات المقالات ضمن نطاق الحد الأدنى والحد الأقصى المشترك.

- توحيد درجات المقال ضمن المجال من [0,10]



➢ توحيد نطاق السمات :Features Scaling

بعد تقسيم البيانات إلى بيانات تدريب وبيانات اختبار قمنا بداية بتوحيد نطاق سمات الدخل بتطبيق StandardScaler .

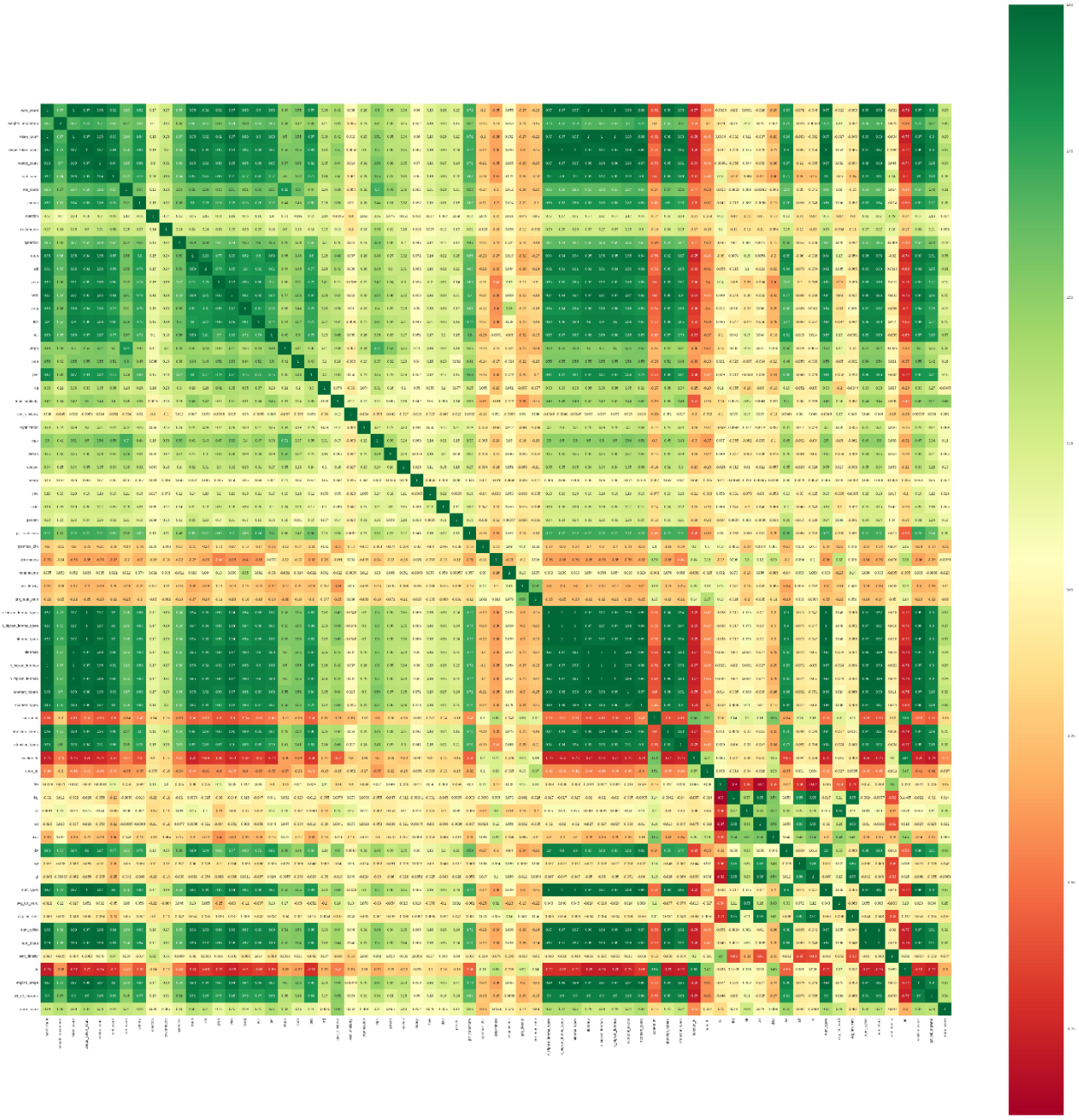
بالتالي يتم تحويل نطاق كل سِمة من السّمات إلى نطاق حول متوسط يساوي الصفر وانحراف معياري يساوي الواحد، أي سيتم تسوية كل قيمة بطرح المتوسط والقسمة على الانحراف المعياري.

السِّمات قبل بتطبيقStandardScaler :

✓ السِّمات بعد تطبيق:StandardScaler

مرحلة اختيار السِّمات وتصوّرها:

- ❖ قمنا بدراسة العلاقات الخطية بين درجات المقال والسِّمات كل منها على حدى، وذلك لمعرفة السِّمات التي لها تأثيراً فعَّال في توقع درجات المقالات.
- ❖ قمنا بعد ذلك باستخدام heatmap لعرض الترابطات بين السِّمات ومدى تأثيرها على سمة درجة المقال.



- ❖ قمنا بعد بذلك بحذف تلك السِّمات بسبب الارتباط الكبير فيما بينها وتلك التي لاتؤثر بشكل فعَّال على توقع درجة المقال.
- ❖ بالتالي قللنا عدد السِّمات من 68 إلى 46 سمة.

مرحلة استخراج سِمات جديدة من أخرى موجودة مسبقا .

بتطبيق خوارزمية Principal Component Analysis (PCA) الإحصائية يتم تحويل السمات المرتبطة خطياً إلى سمات غير مرتبطة خطياً وبالتالي ستساعدنا في:

- ✓ إيجاد علاقة متبادلة بين المتغيرات في البيانات.
- ✓ تفسير وتصور البيانات.
- ✓ تقليل عدد الأبعاد مما يجعل تحليل البيانات أكثر بساطة.

ينتج لدينا فضاء جديد عدد أبعاده أقل مع المحافظة على أكبر تباين للداتا.

بالتالي قللنا عدد السِّمات من 46 إلى 36 سمة.

مرحلة إعادة التدريب:

- ✓ تم تدريب نفس النموذج على السمات ال 36 الناتجة عن تطبيق خوارزمية الPCA

بإسقاط الداتا من 46 بعد ل36 أصبح لدينا خسارة في المعلومات وهناك احتمالين للخسارة:

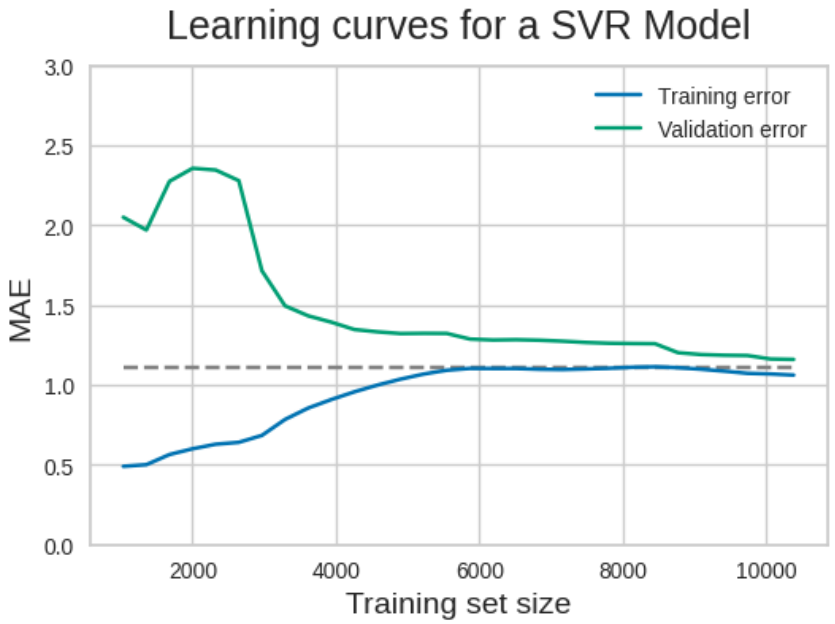
إما أن تكون خسارة للتباين المرتبط بالضجيج وبالتالي سيكون تأثيرها إيجابياً بالنسبة للمتنبأ بدرجة المقال.

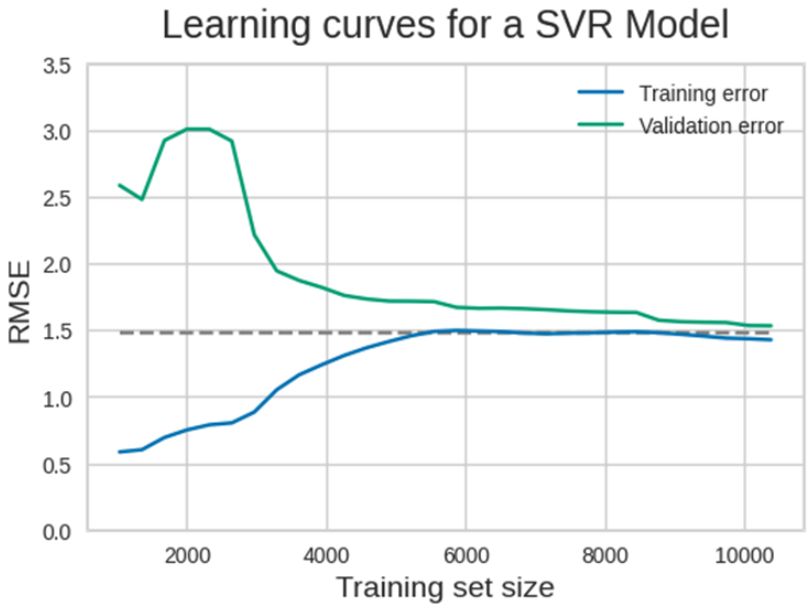
أو أن تكون خسارة لمعلومة أساسية فممن الممكن أن يكون تأثيرها سلبيا

أفضل المتغيرات بعد تطبيق الgridSearsh:

SVR(kernel='rbf',C= 100, epsilon =0.6, gamma =0.004)

في حالتنا الخسارة لم تكن لمعلومة أساسية وذلك لحصولنا على قيم لمعايير التقييم قريبة جداً من بعضها في حال لم نقم بالإسقاط.





Model	RMSE	MAE	QWK
SVR	1.52	1.15	0.7328
SVR + PCA	1.53	1.15	0.7304

النتائج والمقارنة:

من أجل تقييم أداء النظام، نقوم بمقارنته بنظام مفتوح المصدر AES يسمى " Enhanced AI Scoring Engine" (EASE). هذا النظام هو أفضل نظام مفتوح المصدر شارك في مسابقة ASAP ، وقد احتل المرتبة الثالثة بين جميع الفرق المشاركة البالغ عددها 154. يعتمد EASE على ميزات مستخلصة يدويًا وطرق انحدار كذلك. يمكن تصنيف الميزات التي يتم استخلاصها بواسطة EASE إلى أربع فئات:

- طول الميزات
- أجزاء الكلام (POS)
- تقاطع الكلمات مع الجمل التي تركز على موضوع المقال.
- حقيبة الكلمات Bag of n-grams

بعد استخراج الميزات ، تم استخدام خوارزمية الانحدار لبناء نموذج يعتمد على بيانات التدريب.

حيث تم اعتماد طريقتي انحدار المتجه العام (SVR) والإنحدار البايزي الخطي بنسخته المتقدمة(Bayesian linear ridge regression (BLRR كطرق أساسية لبناء النظام.

بالمقارنة مع باقي النماذج المذكورة في الأوراق البحثية التي قمنا بدراستها نجد أن نظامنا قد حصل على أقل نسبة خطأ بالنسبة لمعيار RMSE للتنبؤ بدرجات المقال.

حيث يمثل أول نموذجين في الجدول أدناه التجارب التي قمنا بها:

Model	QWK	RMSE	MAE	R2
SVR	0.732	**1.52**	1.15	0.57
PCA + SVR	0.730	1.53	1.15	0.58
EASE (SVR)	0.699	-	-	-
EASE (BLRR)	0.705			
Doc2vec [1]	0.85	4.43	-	-
SVM [1]	0.75	8.85	-	-

LSTM [1]	0.54	6.8	-	-
BLSTM [1]	0.36	0.5	-	-
Two-layer LSTM [1]	0.46	0.55	-	-
Two-layer BLSTM [1]	0.48	0.52	-	-
LSTM [1] +W2v	0.76	5.39	-	-
BLSTM [1] +W2v	0.85	4.34	-	-
W2v+Two-layer LSTM [1]	0.69	6.02	-	-
W2v+Two-layer BLSTM [1]	0.82	4.79	-	-
SSWES+LSTM [1]	0.94	2.9	-	-
SSWES+BLSTM [1]	0.95	3.21	-	-
Two-layer LSTM [1] +SSWES	0.94	3	-	-
Two-layer BLSTM [1] +SSWES	**0.96**	**2.4**	-	-
LSTM [8]	0.746	-		-
LSTM+attention [8]	0.731	-		-
CNN+LSTM [8]	0.708	-		-
BLSTM [8]	0.699	-		-

الفصل السادس:
الخاتمة

6.1 - الخاتمة:

تم العمل على هذا المشروع وبناء تطبيق يقوم بإعطاء علامة لمقال باللغة الإنكليزية وبذلك نكون قد قدمنا خدمات تعزيز العملية التعليمية حيث تسهل على الأستاذ عملية تقييم المقالات وتساعد الطالب في التحضير للامتحان.

بناءً على ما تكلمنا عنه بالفصول السابقة فقد تم تحقيق جميع المتطلبات وهي:

- تقديم موضوع معين لمناقشته ضمن مقال يكتبه المستخدم.
- تقييم المقال وإعطاء علامة تقريبية.
- حفظ مقالات المستخدم السابقة مع نتيجة التقييم التي حصل عليها.

6.2 – الآفاق المستقبلية:

- نطمح بالمستقبل أن نحسن من دقة النتائج

- كما نطمح لتحسين التطبيق ليصبح قادراً على تقديم ملاحظات للمستخدم لتحسين كتابته.

الفصل السابع:
المراجع

المراجع:

[1] Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks.

[2] Murray, K. W., & Orii, N. (2012). Automatic essay scoring.

[3] Singh, A., & Pant, D. (2019). Automated essay scoring using machine learning.

[4] Phandi, P., Chai, K. M. A. & Ng, H. (2015). Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression.

[5] Song, S., & Zhao, J. (2019). Automated Essay Scoring Using Machine Learning.

[6] Text Summarization An Extractive Method,

[7] Sentence :Sentence-BERT ,Nils Reimers and Iryna Gurevych Ubiquitous Knowledge ,Embeddings using Siamese BERT-Networks ,Processing Lab (UKP-TUDA) Department of Computer Science ,Technische Universitat Darmstadt

[8] Taghipour, K., & Ng, H. T. (2016, November). A neural approach to automated essay scoring.