المجلة الدولية للهندسة وتقنية المعلومات
journal homepage:www.ijeit.misuratau.edu.ly

تعزيز دقة كشف النصوص المزيفة العميقة باستخدام التعلم العميق وتقنيات المعالجة المسبقة

علي محمد حنقة
كلية تقنية المعلومات
a.hanga.pg@it.misuratau.edu.ly

وسام الترجمان
كلية تقنية المعلومات
w.attjroman@it.misuratau.edu.ly

الملخص— **شهدت التطورات الأخيرة في النماذج اللغوية الكبيرة (Large language models - LLMs) تقدمًا كبيراً في توليد النصوص المشابهة للنصوص البشرية. تميزت هذه النماذج بقدرات كبيرة في كتابة محتوى متنوع مثل المقالات الإخبارية والقصص والنصوص العلمية. هذا يؤكد أهمية اكتشاف هذه النصوص لتجنب المخاطر المحتملة كإنتشار الأخبار المزيفة والسرقة الأدبية وضمان سلامة النص في مجالات مختلفة مثل القانون والتعليم والعلوم.**

في بداية بحثنا هذا كان اكبر تحديٍ لنا هو عدم وجود مجموعة بيانات تحتوي على نصوص عربية مزيفة وحقيقية وكانت مجموعة بيانات النصوص الإنجليزية المتوفرة قديمة ومحدودة وتحتاج لتحديث. بناءاً عليه جمعنا نصوصا عربية وانجليزية مكتوبة بشرياً وانتجنا نصوصاً مزيفة من خلال نماذج لغوية كبيرة. قمنا ببناء نموذجين مختلفين لكشف النصوص الانجليزية والعربية المزيفة العميقة. استخدمنا في النموذج الأول تقنيات المعالجة المسبقة وخوارزمية (long short-term memory network - LSTM). وقد حقق النموذج دقة عالية في كشف النصوص الانجليزية بلغت 96% و 56% للنصوص العربية.

ولتحسين دقة كشف النصوص العربية اقترحنا نموذجا آخرا يعتمد على تقنيات المعالجة المسبقة للنصوص ومزيج من LSTM ثنائية الاتجاه (Bidirectional LSTM – BiLSTM) والشبكة العصبية الالتفافية أحادية الأبعاد (1D CNN-). دربنا النموذج وقمنا باختباره على مجموعة بيانات النصوص الانجليزية والعربية المزيفة والحقيقية وتقييمه وفقاً لمعايير الأداء المختلفة حيث أظهرت النتائج تحسنًا كبيراً في دقة الكشف وصلت 86% للنصوص العربية.

الكلمات المفتاحية— **كشف النصوص المزيفة العميقة، النماذج اللغوية الكبيرة (LLMs)، تقنيات المعالجة المسبقة للنصوص، المحولات Transformers، التعلم العميق (BiLSTM , CNN).**

1. المقدمة

يشهد العالم حالياً عصرًا ذهبيًا في توفر كمية ضخمة ومتنوعة من المعلومات حيث يتم كتابة كميات هائلة من المحتوى النصي يوميًا عبر الإنترنت ووسائل التواصل الاجتماعي [1]. في حين يلعب هذا المحتوى دورًا حيويًا في تعزيز الاتصال والتعلم والمشاركة المجتمعية إلا أنه يطرح أيضًا تحديًا كبيرًا متمثلاً في مواجهة انتشار المعلومات المضللة والأخبار المزيفة والتي أصبحت ظاهرة خطيرة تُشكل تهديدًا للمصداقية والنزاهة مع تطور تقنيات إنشاء المحتوى النصي المزيف [2]. يتم استخدام تقنيات الذكاء الاصطناعي المتطورة لإنشاء محتوى نصي يشبه إلى حد كبير النصوص التي يكتبها البشر والذي يتم تحقيقه من خلال استخدام نماذج اللغة الكبيرة (LLMs) التي يتم تدريبها على كميات هائلة من البيانات النصية. يعرض كلا من [4,3] نماذج توليد مختلفة للنصوص المزيفة العميقة. هذه النماذج تتعلم التقاط الأنماط والأساليب اللغوية الموجودة في النصوص البشرية، مما يسمح لها بإنشاء نصوص يصعب تمييزها عن تلك المكتوبة بواسطة البشر [5].

استلمت الورقة بالكامل في 28 ابريل 2024 وروجعت في 10 مايو 2024
وقبلت للنشر في 15 مايو 2024
ونشرت ومتاحة على الشبكة العنكبوتية في 08 أغسطس 2024.

لذا فإن انتشار هذه النصوص يُشكل تحديًا كبيرًا يمكن أن يؤدي الى انعدام المصداقية في البيئة الرقمية فعلى سبيل المثال يمكن أن يستخدم المحتالون هذه التقنيات لإنشاء رسائل بريد إلكتروني احتيالية مزيفة أو رسائل تصيد تهدف إلى خداع المستخدمين وسرقة معلوماتهم الشخصية [6]. كما يمكن استخدامها لكتابة أخبار مزيفة تهدف إلى تشويه سمعة الأشخاص أو التأثير على الرأي العام حول قضايا سياسية أو اجتماعية حساسة [7].

و بالتالي فإن ظهور النصوص المزيفة العميقة يفرض على العالم ضرورة تطوير تقنيات فعالة لكشفها وهذا ما عملنا على تحقيقه في هذه الورقة.

أ. تحديات كشف النصوص المزيفة العميقة

يواجه الكشف عن النصوص المزيفة العميقة العديد من التحديات المهمة التي يجب معالجتها وأهمها هو التشابه الكبير بين النصوص المزيفة العميقة والنصوص البشرية [2,1]. فكما ذكرنا سابقًا فإن نماذج اللغة الكبيرة قادرة على تعلم محاكاة الأنماط اللغوية البشرية بشكل فعال مما ينتج نصوصًا تبدو كأنها مكتوبة بشرياً وهذا يجعل من الصعب على الطرق التقليدية لمعالجة اللغة الطبيعية (Natural language processing – NLP)) التمييز بين النصوص الحقيقية (البشرية) والمزيفة (الألية) [8].

تتميز نماذج اللغة الكبيرة بقدرتها على التعلم المستمر وهذا يجعلها قادرة على توليد نصوص ذات كفاءة لغوية عالية وتتحسن باستمرار مع توفر المزيد من البيانات مما يضيف تحديا آخرا لكشف النصوص التي أنتجتها. بالتالي هذا يتطلب بشكل مستمر ومواكب للتغيرات الحاصلة تطوير تقنيات وأدوات لمواكبة التطور في هذه النماذج [3]. بالإضافة إلى ذلك، يمكن أن تكون النصوص المزيفة العميقة مصممة بطريقة يمكن من خلالها خداع نماذج الكشف القائمة حالياً على سبيل المثال قد يُحاول المهاجمون خداع تقنيات الكشف من خلال انتاج نصوص مزيفة عميقة بطريقة مراوغة والتفافية باستخدام تقنيات مثل التعتيم أو إعادة الصياغة لإخفاء العلامات التي تُشير إلى أنها مزيفة كما ورد في [9]. نقص البيانات اللازمة لتدريب هذه النماذج هي أحد التحديات المهمة و التي ليس من السهل الحصول على كميات كبيرة منها. هذه التحديات تؤكد على أهمية تطوير تقنيات للكشف عن النصوص المزيفة العميقة وتحسين كفاءتها ودقتها حفاظاً على مصداقية البيانات والمعلومات المتداولة عبر الإنترنت.

ب. دور خوارزميات التعلم العميق في كشف النصوص المزيفة العميقة

يعمل العديد من الباحثون والمؤسسات والشركات لحل مشاكل تحديات التزييف العميق في النصوص والمستمرة بقوة و قد تم دراسة وتطوير تقنيات مختلفة في هذا المجال حيث تشير الأبحاث المتنوعة إلى فعالية نماذج التعلم العميق في تصنيف النصوص ومعالجة اللغات الطبيعية وكشف النصوص المزيفة العميقة بما في ذلك استخدام تقنيات التعلم الآلي (ML) وتقنيات التعلم العميق مثل الشبكة العصبية الالتفافية (CNN) وشبكة (LSTM) [2]. تُعتبر خوارزمية CNN واحدة من أهم النماذج العميقة و تتميز بقدرتها على استخراج السمات المحلية والتركيز على النصوص مثل الكلمات والعبارات والتركيبات اللغوية [10]. يمكن دمجها مع خوارزمية LSTM للتعامل مع التسلسلات واكتشاف النصوص المزيفة

بكفاءة عالية. أثبتت شبكة LSTM فعاليتها في العديد من تطبيقات معالجة اللغات الطبيعية مثل تصنيف النصوص والترجمة الآلية والتعرف على الكلام كما أنها تتغلب على مشكلة اختفاء التدفق في الشبكات العصبية التقليدية مما يمكنها من التعامل مع النصوص الطويلة نوعاً ما حيث تستخدم ذاكرة طويلة المدى تسمح لها بتخزين السياق السابق لاستخدامه في عمليات التنبؤ. هذه الميزة مهمة للكشف على النصوص المزيفة العميقة وقد تم تسليط الضوء عليها في [11]. و لأن النصوص تحتوي على روابط معنوية بين الكلمات والجمل في النصوص فهذا يساعد خوارزمية LSTM على التقاط هذه العلاقات وتحديد الأنماط التي تشير إلى أن النص قد تم كتابته بواسطة نموذج لغة آلي وليس بواسطة الإنسان [12]. لتحسين دقة الكشف باستخدام LSTM يمكن إجراء معالجة مسبقة للنصوص على سبيل المثال يمكن تطبيق تقنيات مثل استخراج الميزات اللغوية (كالتركيب النحوي والسمات الدلالية) أو إجراء تحليل للسياق والتبعيات في النص قبل إدخالها إلى النموذج والتي غالبا ما تساعد في تحسين قدرة النموذج على التمييز بين النصوص الحقيقية والمزيفة بدقة أكبر. بالإضافة لذلك يمكن دمج LSTM مع تقنيات أخرى للكشف عن النصوص المزيفة العميقة مثل استخدام نماذج تصنيف إضافية أو تطبيق تقنيات لاستخراج ميزات إضافية من النصوص. يمكن استخدام هذه الإيجابيات في سياق الكشف عن النصوص المزيفة العميقة لتحسين دقة الكشف والتعرف على النصوص المزيفة.

2. الأعمال ذات الصلة

شهدت مهمة اكتشاف النصوص المزيفة العميقة اهتماماً كبيراً في السنوات الأخيرة وذلك نظرًا لكثافة انتشار المحتوى الذي تم كتابته بواسطة نماذج اللغات الكبيرة والتي أصبحت قادرة على محاكاة النصوص المكتوبة بواسطة البشر بشكل دقيق وواقعي. على مر السنوات الماضية قام مجموعة من الباحثين بإجراء العديد من الدراسات وطوروا أدوات وأساليب وتقنيات متعددة باستخدام التعلم الآلي والعميق بهدف كشف النصوص المزيفة العميقة. هذه الجهود ضرورية لمنع انتشارها وكشف التلاعب بها وللحفاظ على مصداقية وجودة البيانات المتداولة عبر الإنترنت بشكل مستمر.

من الدراسات التي تم إجراؤها في هذا المجال تلك التي قام بها دو وفريقه في عام 2022 لتقييم قدرة نموذج اللغة GPT-3 على توليد نصوص ليس من السهل تمييزها عن النصوص المكتوبة بواسطة البشر [13]. أظهرت نتائج الدراسة أن نموذج GPT-3 قادر على إنتاج نصوص ذات جودة عالية تشبه النصوص المكتوبة بواسطة البشر ما شكل تحديًا جديدًا أمام الباحثين يتمثل في تطوير خوارزميات فعالة للكشف عن النصوص المزيفة العميقة. وفي السنة التالية عمل فريق بحث آخر في [1] على تطوير نظام يمكنه كشف النصوص المزيفة العميقة أو التي تم التلاعب بها بشكل فعال في سيناريوهات العالم الحقيقي الذي غالبًا ما تكون فيه النصوص المزيفة العميقة غير قابلة للكشف بمجرد معاينتها من قبل الخبراء. وقد أجرى الباحثون اختباراتهم في ذلك العمل على مجموعة كبيرة من البيانات النصية الانجليزية قاموا بجمعها من عدة مصادر شملت وسائل التواصل الاجتماعي والمقالات الإخبارية ومنصات الويب. أظهرت النتائج فعالية نهجهم ودقته مما يشير إلى إمكانية استخدام وتطوير في مكافحة التهديد المتزايد والمتغير للنص المزيف العميق. ظهرت طرق متنوعة أخرى تعتمد إحداها على تحليل خصائص النصوص المشتقة من النماذج اللغوية الكبيرة مثل توزيعات n-gram وكثافة الكلمات والتي بناءاً على النتائج يتم مقارنة هذه الخصائص بنصوص مكتوبة بواسطة البشر للكشف عن الاختلافات والتناقضات [14].

فيالدراسات [16,15] طور الباحثون أداة لاختبار قدرة المستخدمين على اكتشاف النصوص المولدة آليًا وتمييزها عن تلك المكتوبة بشريا فهي تطلب من المستخدم تحديد ما إذا كانت الآلة قد كتبت جزءًا من النص معين أو النص بأكمله وفي حالة الإجابة بنعم على المستخدم تحديد نقطة أو نقاط الانتقال في النص من الإنسان إلى الآلة و بالرغم من أن النتائج أظهرت أن عدداً قليلاً جدًا من المشاركين نجحوا في تحديد الفروق الدقيقة بين النصوص البشرية والنصوص المُولَّدة آليًا بشكل صحيح إلا أن هذه الأداة قد ساهمت في جمع مجموعة لا بأس بها من البيانات التعليقية والمؤشرات التي يحتاجها الباحثون في هذا المجال. و في دراسة حديثة [17] أجريت في 2023 أظهر الباحثون أن LLMs قادرة على توليد نصوص يصعب تمييزها عن النصوص المكتوبة من قبل البشر بواسطة الطرق العادية واقترحوا طريقة تعتمد على تحليل خصائص اللغة مثل النحو والدلالات اللغوية. وقد سلط الباحثون الضوء على ضرورة وجود أدوات مساعدة لطرق تمييز النصوص المزيفة. أحد أهم هذه الأدوات تعتمد على المعالجة المسبقة للبيانات وأيضاً أظهرت الدراسة [18] أن تقنيات المعالجة المسبقة مثل الترميز، وإزالة كلمات التوقف، والقطع، والتطبيع تؤثر بشكل كبير وملحوظ على أداء نماذج تصنيف النص مما يحسن دقتها ويقلل التشويش في البيانات. كما تناولت عدة دراسات أخرى عوامل مختلفة مؤثرة على فعالية خوارزميات الكشف مثل تنوع مجموعة البيانات وقوة النموذج. من بين تلك الدراسات نشير إلى [8،2].

تم استخدام تقنيات التعلم العميق وتحديدًا شبكات LSTM على نطاق واسع في العديد من مهام معالجة اللغات الطبيعية نظرًا لقدرتها على التقاط التبعيات طويلة المدى [11]. وفي عام 1997 قدم المؤلفون في [19] شرحًا شاملاً لبنية LSTM. و بينوا أنه من الممكن استخدام هذه الشبكات على نطاق واسع في تصنيف النصوص وتحليل المشاعر والترجمة الآلية ومهام البرمجة اللغوية العصبية. كما توجد طريقة تجمع بين تقنية Multi-LSTM وتقنية blockchain قدمها تشان [20] لاكتشاف التزييف العميق في الوسائط المتعددة. و على الرغم من أن ذلك العمل ركز بشكل رئيسي على محتوى الوسائط المتعددة إلا أن استخدام شبكات LSTM يوضح قدرتها على التقاط التبعيات التسلسلية واكتشاف المعلومات التي تم التلاعب بها مما يجعلها مناسبة لمهام NLP المعقدة مثل كشف النصوص المزيفة العميقة. استخدم روي [21] نماذج الشبكات العصبية المتكررة LSTMs و Bi-LSTMs ووحدة التكرار المبوب (Gated Recurrent Unit - GRU) لمكافحة هجمات التصيد الاحتيالي المتزايدة واكتشاف عناوين URL الضارة. و في دراسة أخرى في عام 2023 قام بها الفريق ذاته [22] تم تطوير نموذج للكشف عن خطاب الكراهية باستخدام توجيهات LSTM و تقنية TF-IDF. هذا النموذج يتميز بدقته في تصنيف مشاعر الكراهية مقارنة بالنماذج الأخرى المتاحة.

للاستفادة من قوة كلاً من CNN وLSTM يمكن بناء نهج يضم ليهما قادر على التقاط الخصائص المكانية للبيانات النصية بشكل مميز وفعال من خلال شبكة CNN بينما تلتقط LSTM العلاقات الزمنية. تقدم [23] نموذجا مبتكرا وهو عبارة عن نموذج هجين للتعلم العميق يدمج بين CNN وLSTM بهدف اكتشاف الأخبار الكاذبة. أظهر هذا النموذج بحسب الدراسة دقة عالية في الأداء. نستنتج من هذا أن التكامل بين CNN و LSTM يوفر نموذج قوي وشامل لاكتشاف النصوص المزيفة العميقة. مما يعكس القدرة المتزايدة للتقنيات القائمة على التعلم العميق في معالجة اللغة الطبيعية وتحليل المحتوى النصي. وبالرغم من فعالية LSTM وكفائتها إلا أنها قد تواجه قيودًا في التقاط العلاقات المحلية والطويلة الأجل في المدخلات وفي دراسة سابقة خلال عام 2019 طور المؤلفون نموذج قائم على شبكة BiLSTM للتمييز بين الحسابات الآلية والحسابات البشرية على تويتر دون الحاجة إلى معرفة مسبقة أو ميزات مصممة يدويًا فقد أظهرت تجاربهم أن نظامهم المقترح حقق أداءً معقولاً مقارنةً بأنظمة كشف الحسابات الآلية المماثلة كما أكدوا على قدرة نموذجهم على كشف البريد الإلكتروني الاحتيالي وصفحات الويب والرسائل القصيرة [24].

إقترح باحثون من Google بنية مطورة تسمى المحول (Transformer) والتي تعتمد بشكل أساسي على آلية الانتباه متعدد الرؤوس (Multi-Head Attention) تم تقديمها لأول مرة في [25] سنة 2017 بعنوان "الانتباه هو كل ما تحتاجه". يعمل المحول على تحويل النص إلى تمثيلات رقمية تُسمى (Tokens) والتي تحول كل رمز مميز إلى متجه عبر البحث في جدول تضمين الكلمات (Word Embedding Table) في كل طبقة ليتم لاحقا وضع كل رمز مميز في سياقه ضمن نطاق نافذة السياق (Context Window) مع الرموز المميزة الأخرى عبر آلية الانتباه متعدد الرؤوس والتي بدورها تسمح بتضخيم إشارة الرموز المميزة الرئيسية وتقليل أهمية الرموز المميزة الأقل أهمية مما يُساعد على فهم العلاقات بين الكلمات والجمل في النص. هذه الالية تجعل المحول أداة قوية لكشف النصوص المزيفة العميقة وخصوصاً الطويلة. يمكن استخدام طبقة (Transformer block) والتي تعتبر المكون الأساسي في بنية المحول كأحد مكونات نموذج كشف النصوص المزيفة العميقة حيث تُساعد هذه الطبقة على فهم العلاقات بين الكلمات والجمل في النصوص مما يُعزّز قدرة النموذج على تمييز التناقضات والأنماط غير الطبيعية التي قد تشير إلى محتوى مكتوب آلياً.

ومن الطرق المهمة المستخدمة في كشف النصوص المزيفة العميقة هي طرق التضمين التقليدية والتي تعتمد على استنتاج واستخراج تمثيلات للكلمات استناداً إلى سياقها في النص. وعلى الرغم من قدرة هذه الطريقة على التقاط بعض معاني الكلمات الفردية إلا أنها قد لا تمتلك القدرة على استيعاب الفروق الدقيقة في اللغة مما قد يؤدي إلى تراجع دقة الكشف عن

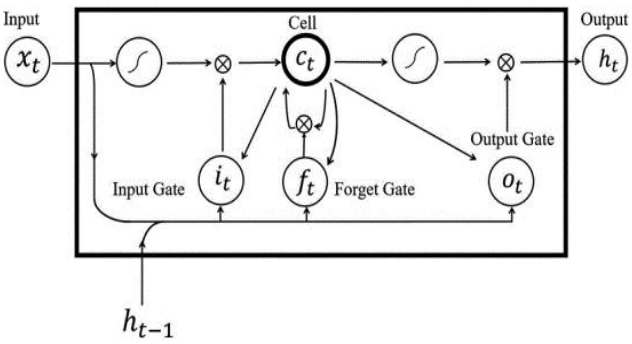
النصوص المزيفة العميقة. لتفادي هذه المشكلة تقدم [26] نموذجا لغويا بديلا يسمى (Bidirectional Encoder Representation of Transformers - BERT) يعتمد في بناءه على معمارية المحولات حيث يقوم ببناء تمثيلات عالية الجودة للكلمات مما يعزز أداء مهام معالجة اللغات الطبيعية بشكل عام .

توفر هذه الدراسات نظرة عامة على أبرز التحديات والفرص والتطورات في مجال كشف النصوص المزيفة العميقة. من خلال معالجة هذه المشاكل يهدف الباحثون إلى تطوير أنظمة كشف أكثر قوة وموثوقية قادرة على التقليل من انتشار المعلومات المزيفة.

3. الخوارزميات المستخدمة

أ. الذاكرة الطويلة قصيرة المدى

تتميز بقدرتها على معالجة سلاسل كاملة من البيانات من خلال وصلات التغذية الراجعة التي تمتلكها، ومن المكن تطبيقها على مهام متنوعة مثل معالجة اللغات الطبيعية وتحليل التسلسلات الزمنية والترجمة الآلية، وغيرها [19]. يمكنها أن تتذكر المعلومات على فترات زمنية طويلة من خلال وحدة خلية الذاكرة والتي تعمل كوحدة ذاكرة تستخدم سلسلة من العمليات تعمل على تحديث حالتها من خلال سلسلة من العمليات التي تُنظمها ثلاث بوابات رئيسية هي: (بوابة الإدخال) - (بوابة النسيان) - (بوابة الإخراج) تسمح هذه البوابات التي يتم تحديد معلماتها بواسطة أوزان قابلة للتعلم بقراءة المعلومات وكتابتها ونسيانها بشكل انتقائي في كل خطوة زمنية. تتحكم بوابة الإدخال في كمية المعلومات الجديدة التي تتم إضافتها إلى الحالة الداخلية للخلية بينما تتحكم بوابة النسيان في المعلومات التي يتم التخلص منها حيث تحدد بوابة الإخراج مقدار المعلومات التي يتم تعريضها للطبقة التالية التي تستخدمها للتنبؤات. تقوم هذه البوابات بتنظيم تدفق المعلومات إلى داخل الوحدة وخارجها، وهذا ما يجعلها مثالية لمعالجة البيانات التسلسلية والتنبؤ بها تم تطويرLSTM للتعامل مع مشكلة التلاشي المتدرج، والمتمثلة بفقدان المعلومات طويلة المدى في تسلسل البيانات وبفضل قدرتها على ضبط بواباتها بناءً على تسلسل الإدخال تتمكن الـ LSTM من التقاط التبعيات طويلة المدى بشكل فعال أثناء التدريب [27]. المدخلات تتكون من جزئين: حالة الإدخال في الوقت الحالي (x_t) t والتي تمثل المعلومات الجديدة الواردة إلى الشبكة في هذه اللحظة. حالة الإخراج في الوقت السابق (h_{t-1}) $t-1$ تمثل المعلومات التي تم إخراجها من الشبكة في الزمن السابق *(t-1)* [28]. الشكل1 يوضح بنية وحدة LSTM.



الشكل1 بنية وحدة LSTM

تسيطر بوابة الإدخال على كمية المعلومات الجديدة التي ستخزن في خلية الذاكرة حيث تأخذ في الاعتبار الإدخال الحالي (x_t) والحالة المخفية للخطوة الزمنية السابقة (h_{t-1})، وتنتج قيمة بين 0 و 1 لتحديد أهمية المعلومات الجديدة. يتم تمثيلها بالمعادلات (1) و (2) التالية:

$$i_t = \sigma (W [x_t , h_{t-1} , C_{t-1}] + b_i) \quad (1)$$

$$C_t = f_t \cdot C_{t-1} + i \cdot tanh (W [x_t , h_{t-1} , C_{t-1}] + b_c) \quad (2)$$

تتحكم بوابة النسيان في كمية المعلومات التي سيتم إزالتها من خلية الذاكرة. تأخذ في الاعتبار المدخل الحالي (x_t) والحالة المخفية السابقة (h_{t-1}) ومررها إلى وظيفة تنشيط سيجما، والتي تنتج قيمة بين 0 و1، حيث يعني 0 نسيان و1 يعني الاحتفاظ. كماهو موضح في معادلة (3) التالية:

$$f_t = \sigma (W_f [h_{t-1} \cdot x_t] + b_f) \quad (3)$$

تعمل بوابة النسيان وبوابة الإدخال على تحديث حالة الخلية. يتم ضرب حالة الخلية للحالة السابقة بمخرجات بوابة النسيان. ثم يتم جمع مخرجات هذه الحالة مع مخرجات بوابة الإدخال. يتم بعد ذلك استخدام هذه القيمة لحساب الحالة المخفية في بوابة الإخراج. كماهو موضح في معادلة (4) التالية:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \check{C}_t \quad (4)$$

وأخيرا، من خلال بوابة الإخراج يتم الحصول على نتيجة الإخراج .يتم تمثيلها بالمعادلات (5) و (6) التالية:

$$o_t = \sigma (W_o \cdot [h_{t-1} , x_t] + b_o) \quad (5)$$

$$h_t = o_t \cdot tanh (C_t) \quad (6)$$

تقوم شبكة LSTM بمعالجة تسلسل النص كلمة كلمة في كل خطوة للكشف عن النص المزيف العميق. تعمل وحدة LSTM على تحليل الكلمة الحالية جنبًا إلى جنب مع السياق الذي توفره الحالة المخفية السابقة. يتيح ذلك للشبكة التقاط العلاقات بين الكلمات وتحديد الأنماط التي قد تشير إلى خصائص نصية مزيفة عميقة. من خلال تحليل هذه الأنماط عبر تسلسل النص بأكمله.

ولكن بالرغم من أن LSTM تغلبت على مشاكل الذاكرة قصيرة المدى من خلال إدخال آليات البوابات الداخلية التي تنظم تدفق المعلومات، إلا أنها تقوم بنقل المعلومات في اتجاه واحد فقط من الأمام إلى الخلف. لاحقاً تم معالجة هذه المشكلة من خلال اقتراح شبكة LSTM ثنائية الاتجاه (BiLSTM) والتي تجمع بين شبكتين LSTM تعملان بشكل متوازي لتمرير نفس تسلسل الإدخال إلى LSTM الأمامي والخلفي ودمج الطبقات المخفية لكل منهما معًا بطبقة الإخراج للتنبؤ. أثبتت شبكات BiLSTM نجاحها في مجموعة متنوعة من التطبيقات مثل الترجمة الآلية، ومعالجة اللغات الطبيعية والتنبؤ بالامراض [30]. تعمل الشبكة الأمامية بمعالجة التسلسل في الاتجاه الأمامي، وتعالج الشبكة الخلفية التسلسل في الاتجاه العكسي، ومن ثم يتم دمج مخرجاتهما معاً. تتميز شبكة BiLSTM باستخلاص الارتباطات بين الكلمات في الجملة ولتعزيز قدرة النموذج على فهم توجهات الكلمات وتحليل السياق [31]. من الممكن توجيه نموذج BiLSTM لمعالجة التسلسل واستخلاص المعلومات منه في كلا الاتجاهين ومنها يتم دمج هذه المعلومات معًا للوصول إلى النتيجة النهائية. هذا النهج يعطي النموذج قدرة أكبر على التعامل مع الارتباطات الثنائية وتحليل السياق في المهام المختلفة.

ب. الشبكة العصبية الالتفافية

يعد نموذج CNN خيارًا مثاليًا ومناسبًا تمامًا بسبب قدرته على استخراج الميزات المهمة من البيانات النصية [33]. يتكون من مجموعة متتالية من الطبقات الالتفافية، تقوم بتجميع بيانات الإدخال باستخدام المرشحات القابلة للتعلم، ومن ثم تتبعها وظائف التنشيط غير الخطية، وعمليات التجميع للاختزال والطبقات الكثيفة. تُستخدم الطبقات الالتفافية لاستخراج السمات المهمة من تضمينات الإدخال بينما يقوم الحد الأقصى للتجميع بتقليل الأبعاد والتقاط الميزات البارزة. وفي النهاية، تُستخدم الطبقات الكثيفة للتصنيف، وذلك باستخدام السمات المستخرجة. يتميز هذا النموذج بقدرته على تعلم التمثيلات المميزة للميزات من النص الخام بشكل تلقائي دون الحاجة إلى تدخل بشري لهندسة الميزات وبالتالي يمكن للنموذج التكيف مع مختلف أنواع البيانات النصية وتعلمها من مجموعات بيانات واسعة النطاق دون تدخل بشري في هذا [27] قدم الباحثون نموذج CNN عميق يتعلم بشكل فعال السمات المميزة لاكتشاف المعلومات الخاطئة. الا أن هذا النموذج يعاني من بعض التأثيرات السلبية للتركيب الزائد مما يتطلب مدة أطول للتدريب، إلا أنه يمكن أن يؤدي أداء أفضل في مهام كشف النصوص المزيفة العميقة من خلال القدرة على اكتساب أنماط معقدة تدريجيًا داخل النص.

ج. الشبكة العصبية الالتفافية أحادية البعد

تعتبر (1D-CNN) أكثر ملاءمة للتعامل مع بيانات الإدخال أحادية الأبعاد، مثل البيانات الزمنية والإشارات الطبية الحيوية [29]. تستخدم

لمعالجة البيانات التي يتم تمثيلها على شكل تسلسلات أو سلاسل من البيانات.
المعادلة (7) تُعبرعن عملية التفاف نواة (W) بإشارة إدخال أحادية البعد (S).

$$(S * W)_n = \sum_{i=1}^{|w|} W_i S(i + n - 1) \qquad (7)$$

حيث تقوم عملية الالتفاف بإخراج خريطة الميزات.
تبدأ عملية التفاف نواة التفاف W بإشارة إدخال S من الموقع الأول (n=1).
في كل موقع n يتم ضرب عناصر نواة التفاف W بعناصر إشارة الإدخال S من الموقع $(i + n - 1)$، حيث i يتراوح من 1 إلى طول نواة التفاف W.
يتم جمع نتائج الضرب للحصول على قيمة واحدة في (feature map) في الموقع n.
تتكرر عملية التفاف نواة التفاف W بإشارة إدخال S على جميع المواقع n، مما ينتج عنه خريطة خصائص كاملة.

ولاحقاً يتم تمثيل مصفوفة الإخراج بالمعادلة (8):

$$O'_n = (S_{|w(i,j)} * w (i , j))_n \qquad (8)$$

حيث تصف هذه المعادلة عن مصفوفة الناتج O'_n *(output matrix)* في الموقع n.

د. المحول

المحول هو نموذج يستخدم تقنية الانتباه لفهم البيانات النصية بشكل أفضل حيث يعتمد عليها لتحديد الأجزاء المهمة في البيانات النصية عن طريق منحها وزناً أكبر لتسلط الضوء على الأجزاء المهمة. يتم استخدام الانتباه متعدد الرؤوس لالتقاط كمية أكبر من المعلومات مقارنة بالانتباه ذي الرأس الواحد. من خلال استخدام عدة رؤوس لحساب الانتباه من منظورات مختلفة للبيانات، حيث تتمتع آلية الانتباه متعددة الرؤوس بقدرة على تحديد الأنماط أو الميزات الرئيسية المختلفة في نفس الوقت ومن خلال تسهيل الهيكل الخاص بالمحول المعالجة الموازية ليتم التعلم بشكل اسرع ومعالجة تتم بسلاسة مع كميات كبيرة من البيانات. يعتبر المحول نموذجًا عصبيًا قويًا يُمكن دمجه مع تقنيات أخرى مثل BiLSTM و CNN يمكن بناء نماذج أكثر فعالة وقادرة على كشف النصوص المزيفة العميقة.

4. تجميع البيانات

هدفنا من بداية هذا العمل هو تطوير نظام فعال لكشف النصوص المزيفة العميقة، ولتحقيق ذلك، يجب أن يتوفر لدينا مجموعة بيانات تتناول موضوعات متنوعة تحتوي مزيجاً متنوعاً من النصوص المزيفة المولدة بواسطة نماذج اللغة الكبيرة LLMs المختلفة، وايضا نصوصاً حقيقية مكتوبة بواسطة البشر والتي تم جمعها من مصادر موثوقة مختلفة. على أن يكون حجم مجموعة البيانات كبيرًا قدر الإمكان لضمان تدريب وتقييم النماذج المقترحة بدقة. جمع مجموعة البيانات ومعالجتها مسبقًا كانت هي المرحلة الأولية والتي قمنا خلالها بجمع مجموعتي بيانات منفصلتين الأولى باللغة الانجليزية والثانية باللغة العربية وحفضها بصيغة CSV، لاستخدامها لتدريب وتقييم النماذج في كشف النصوص المزيفة العميقة باللغتين العربية والإنجليزية. تتكون المجموعة الأولى من بيانات النصوص الإنجليزية المزيفة والحقيقية والتي تضم 60365 نص قصير ومقالات طويلة وأخبار ومعلومات عامة والتي مجموعة متنوعة من المجالات والمحتوى النصي. أما المجموعة تتكون مجموعة بيانات النصوص العربية المزيفة والحقيقية فتضم 4200 نص قصير ومقالات طويلة وأخبار ومعلومات عامة. قمنا بجمع البيانات المزيفة من خلال استخدام نماذج توليد اللغة التالية GPT-4 و GPT-3.5-Turbo وGemini وLlama وClaude3. وهكذا تمكنا من الحصول على نصوص مزيفة باللغتين العربية وحفظ كل منهما في مجموعة بيانات منفصلة.. بالإضافة إلى النصوص التي تم إنشاؤها تتضمن مجموعات البيانات أيضًا نصوصًا ومقالات وأخبار كتبها البشر والتي تم الحصول عليها من خلال جمعها من مصادر موثوقة مثل قناة الجزيرة، بي بي سي نيوز، سي إن إن، سكاي نيوز، العربية، العربي، الغارديان، صحيفة الحياة، صحيفة الشرق الأوسط، صحيفة الرياض والتي تشتهر بمصداقيتها وتوفر مجموعة متنوعة من النصوص. وجمعنا أيضاً من بعض الكتب والبعض الاخر تم كتابته من قبلنا، مما يضمن إدراج محتوى موثوق في مجموعة البيانات. تم حفظ مجموعات البيانات في ملفات منفصلة من نوع CSV وصنفناها على أنها نصوص "مزيفة عميقة" واشرنا اليها بالرقم 1 و "حقيقية بشرية" واشرنا اليها بالرقم 0. تعرض الجدول 1 و 2 تفاصيل عدد النصوص واللغة ومصدرها لمجموعتي البيانات المستخدمة في هذا البحث.

جدول1: توزيع أعداد النصوص المولدة

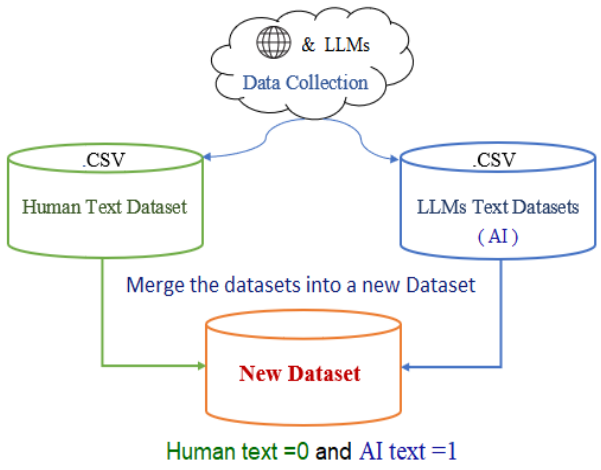
التصنيف	GPT-4	GPT-3.5 Turbo	Gemini	Claude	Llama
English	1803	10496	9682	7560	1967
Arabic	0	857	931	347	0

جدول2: توزيع أعداد النصوص البشرية

التصنيف	BBC	CNN	Sky news	AL Arabia	Aljazeera	Elaraby
English	1928	2098	1422	-----	-----	-----
Arabic	114	178	42	203	183	55

التصنيف	The guardian	alhayat	alriyadh	eurone ws	RT Arabic	Another
English	1542	-----	-----	1986	-----	19881
Arabic	-----	39	287	-----	57	907

ولتجهيز مجموعتي البيانات للتحليل والمعالجة المسبقة تم دمج كلا من مجموعة البيانات البشرية والمولدة بواسطة LLMs في مجموعة تضم النصوص الإنجليزية ومجموعة تضم النصوص العربية كما في الشكل2.



شكل2. مجموعة بيانات النصوص البشرية ومجموعات بيانات النصوص الآلية

5. المعالجة المسبقة للبيانات

قبل استخدام البيانات كمدخل لنماذجنا من المهم ان تمر بمراحل من المعالجة والتي تعتبر ضرورية في أي مهمة من مهام معالجة اللغات الطبيعية (NLP)، والتي تتضمن تحويل بيانات النص الخام إلى تنسيق يمكن للخوارزميات فهمه وتحليله بحيث تتكون هذه المرحلة من عدة خطوات أساسية مهمة لمعالجة وإعداد البيانات النصية منها استخراج الميزات، والترميز، والتجذير، والتوجيه. يتم استخراج الميزات بناءً على تحديد سمات محددة من البيانات النصية الخام في البرمجة اللغوية العصبية وغالباً ماتتعلق هذه الميزات بمجموعات الكلمات داخل النص أو الحروف فالترميز يعتبر جزءًا مهماً وأساسيًا من عملية استخراج الميزات وبعد الترميز وفي خطوة لاحقة يتم توجيه هذه البيانات من خلال تحويل النص إلى تمثيل متجهي لتدريب النماذج. هناك طرق مختلفة تعمل على تحويل البيانات النصية إلى متجهات رقمية مثل مصطلح تردد الكلمة - تردد

الوثيقة المعكوس (Term Frequency-Inverse Document Frequency TF- IDF)، وكيس الكلمات (Bag of Words - BoW)، كما تم عرضها في [18]. في بعض الحالات، يمكن أن تكون هناك كلمات محددة تستخدم بشكل متكرر ولكنها قد لا توفر معلومات مفيدة يمكن الاستفادة منها وهو ما يسمى بكلمات التوقف مثل حروف الجر والضمائر والتي من الممكن أن يساعد إزالة هذه الكلمات في تقليل أبعاد البيانات وزيادة الكفاءة الحسابية [34]. يتم استخدام تقنية (Stemming) بهدف إختزال الكلمات إلى جذورها الأساسية في اللغات العربية والإنجليزية، وهذا يساعد في تحسين دقة تصنيف النصوص وتقليل أبعاد البيانات وتحسين أداء البحث عن المعلومات.

بشكل عام، تتطلب منا معالجة مجموعتي البيانات مجموعة من الخطوات المهمة لتهيئة وتجهيز البيانات قبل إدخالها إلى النماذج المقترحة. تنظيف النصوص وتهيئتها من أهم الخطوات، والذي يتطلب منا بشكل عام إزالة الزخم اللغوي الزائد والرموز الزائدة، لنضمن استخدامنا لبيانات خالية من الشوائب والعيوب. قما بازالة الأحرف غير الضرورية والرموز الخاصة وبعض الحروف الزائدة مثل الفواصل والنقاط الزائدة و "ال" و "اء" و "ة" للنصوص العربية، و إزالة علامات الترقيم، والرموز غير الضرورية، وإزالة بعض البادئات واللواحق الشائعة مثل "un-" و "-ing"، وإزالة الكلمات الشائعة غير المفيدة مثل ("the" و "a") من النصوص لتنظيف البيانات وتسهيل المعالجة اللاحقة. قد تحتوي البيانات على أرقام، ولكن تحويلها إلى نص يساعد في تجنب الخلط بينها وبين الكلمات الأخرى على سبيل المثال، "123" تتحول إلى "مئة وثلاثة وعشرون"، مما يجعل التعامل مع البيانات أكثر فهماً ودقة. ولتقليل التداخل تمت إزالة الرموز التعبيرية والنصوص الفارغة وإزالة الهمزات وعلامات التشكيل، وفصل الكلمات المتصلة من النصوص العربية، وتحويل الحروف إلى صيغة الحروف الصغيرة في النصوص الانجليزية لتجنب التمييز بين الحالات الكبيرة والصغيرة. بالإضافة إلى ذلك، استبدلنا روابط موقع الويب بسلسلة فارغة، وقمنا أيضًا بتطبيع النصوص، وإزالة الأحرف غير العربية، وإزالة علامات التشكيل مثل الهمزة والتشديد وإزالة علامات الإعراب من الكلمة مثل الضمة والفتحة. في النصوص العربية، و استبدلنا القيم المفقودة بقيم "غير محددة"، واستخدامنا التطبيقات اللغوية المتقدمة مثل قواعد التجذير لإرجاع الكلمات إلى جذورها الأساسية على سبيل المثال، كلمة "السلام" و "سلامة" تشتركان في الجذر "سلم"، و "running" إلى "run" و "playing" إلى "play" في النصوص الانجليزية. تم تقسيم كلاً من مجموعة البيانات الانجليزية ومجموعة البيانات العربية المزيفة إلى 70% لتدريب النماذج كل منها لوحده، و15% للتحقق من الصحة لضبط المعلمات وتحديد أفضل النماذج، و15% لاختبار وتقييم أداء النماذج النهائية. تتكون كل مجموعة بيانات من عمودين: النص، والتسمية (0- مزيف (AI)، 1- حقيقي (Human)).

تم تطبيق التجزيء لتحويل الكلمات إلى وحدات دقيقة (Tokenz) هذا يساعد في تحسين جودة التحليل وتقليل التشتت في البيانات. بالاضافة لترميز النصوص لتحويل الكلمات إلى أرقام من خلال استخدام تقنيات مثل ترميز الكلمات (word embedding) لتمثيل الكلمات كنقاط ليتم تغذيتها إلى النموذج. من خلال تنفيذ هذه الخطوات، يتم الحصول على نصوص نظيفة وجاهزة للمعالجة باستخدام نماذجنا.

6. النماذج المقترحة

عملنا في هذه الدراسة على تطوير نموذجين مختلفين لكشف النصوص الإنجليزية والعربية المزيفة العميقة. قمنا بتدريب وتقييم واختبار كل نموذج بشكل منفصل بحيث يقوم كل نموذج بالتنبؤ على أساس أنها نصوص مزيفة (مولدة بواسطة نموذج لغة) أو حقيقية (مكتوبة بشرياً). يتكون كل نموذج من عدة مراحل والتي تم تفصيلها في كل نموذج كالتالي:

أ. النموذج الأول المقترح

استخدمنا نموذج LSTM كنموذج أول لكشف النصوص الانجليزية والعربية المزيفة العميقة. أضفنا طبقات وتعديلات كمحاولة لتحسين دقة النموذج وزيادة قدرته على تمثيل السياق بشكل أفضل. بعد الادخال بدأنا بتحويل النصوص إلى تمثيلات رقمية باستخدام مصفوفة الـ Embedding واستخدمنا طبقة LSTM لفهم السياق والترتيب الزمني للكلمات وفي المرحلة الثانية أضفنا طبقات من LSTM لزيادة قدرة التعبير والتمثيل العميق للنموذج والتي تم تمثيلها بالمعادلة (9) التالية:

$$h^{L}_{t} = LSTM\,(x_t, h^{L}_{t-1}) \qquad (9)$$

حيث (h^{L}_{t}) هو الحالة المخفية في الطبقة الإضافية في الزمن (t) و (h^{L}_{t-1}) هو الحالة المخفية في الطبقة الإضافية في الزمن (t-1).

، وتطبيق تقنيات التنظيم مثل التساقط (*Dropout*) على الحالة المخفية في طبقة LSTM. والذي يساهم في تقليل التشوش وتحسين قدرة النموذج على التعميم. يمكن تمثيل ذلك بالمعادلة (10) التالية:

$$h^{\mathrm{drop}}_{t} = Dropout\,(h_t, P) \qquad (10)$$

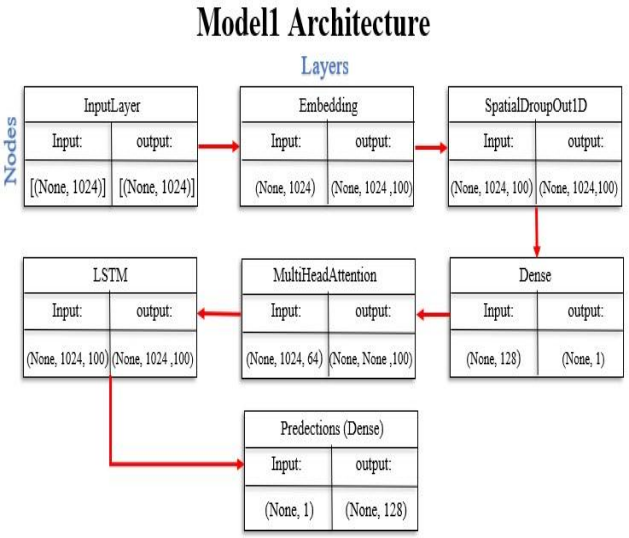
حيث (h^{drop}_{t}) هو الحالة المخفية بعد تطبيق التساقط، (h_t) هو الحالة المخفية قبل تطبيق التساقط، و (P) هو احتمال التساقط.

كما قمنا بإضافة طبقات التعامل مع الانتباه (*Attention*) لتحسين قدرة تركيز النموذج على الجوانب الأكثر أهمية والذي يحسن قدرته على تمثيل العلاقات البينية بين الكلمات. تم تمثيل ذلك بالمعادلة (11) التالية:

$$a_t = Attention\,(h^{L}_{t}, h_s) \qquad (11)$$

حيث (a_t) هو النسبة المرتبطة بالتركيز على الكلمة المطلوبة في السياق، (h^{L}_{t}) هو الحالة المخفية في الطبقة الإضافية، و (h_s) هو مجموعة حالات المخفية للسياق.

مما لاشك فيه أن المعالجة المسبقة للبيانات تساعد في تحسين دقة النموذج لما لها من دور كبيرة في تجهيز البيانات بشكل صحيح للتدريب وتحسين قدرة النموذج على التعلم من البيانات والذي بدوره يمكن أن يحقق دقة عالية في كشف النصوص المزيفة.



شكل3. بنية النموذج الأول

أ. النموذج الثاني المقترح

محاولةً منا للرفع من دقة الكشف قمنا في النموذج الثاني ببناء مجموعة متنوعة من الطبقات والتقنيات والمتمثلة في إستخدمنا أولاً لطبقة التضمين والتي تقوم بتحويل النصوص إلى تمثيلات رقمية قابلة للتعامل معها بواسطة النموذج. تأخذ هذه الطبقة أولا متغير الإدخال (inputs) الذي يكون شكله (sequence_length) وتقوم بتحويله إلى تمثيلات متطابقة في الأبعاد و باستخدام وظيفة التضمين يتم تحديد أقصى عدد للميزات (max_features) وأبعاد التضمين (embedding_dim) في هذه الطبقة وهو يعبر عن حجم المصفوفة الناتجة.

والتي تليها ثاني الطبقات وهي طبقة Bidirectional LSTM لاستخدامها في استخلاص المعلومات السياقية من النصوص لتحسين قدرة النموذج على التعامل مع السياقات المعقدة حيث تأخذ التضمين كإدخال وتعيد تسلسل النتائج للاستفادة من المعلومات في السياق السابق واللاحق.

و أهم الطبقات هي الطبقة الثالثة Depth Transformer Block :

وهي من الطبقات الأساسية المستخدمة في النماذج التحويلية (Transformers) والتي تعتبر فعالة في استخراج العلاقات البينية في البيانات التسلسلية وتتكون هذه الطبقة من عدة مكونات، بما في ذلك طبقة الانتباه المتعددة الرؤوس (MultiHead Attention) الذي يستخدم لاستخراج العلاقات البينية بين الكلمات في النصوص حيث يتمثل الإخراج في تمثيلات تركيز متعددة للنصوص والذي يتم تمثيله بالمعادلة (12):

$$attn\_output = MultiHead\ Attention\ (\ inputs\ ,\ inputs\ ) \quad (12)$$

والطبقة الكاملة المتابعة (Feed-Forward Neural Network) تستخدم لتحسين التمثيل العميق للبيانات بتطبيق عمليات التحويل الغير خطية لاستخلاص الميزات الهامة. تُمثل بالمعادلة (13):

$$affn\_output=FeedForwardNetwork(out1) \quad (13)$$

وطبقات التعامل مع الانتباه (Layer Normalization) والتي تقوم بتحسين استقرار الإخراج من طبقة الانتباه المتعددة الرؤوس من خلال العمل على إعادة تسوية توزيع البيانات لتحسين قابلية التدريب وتُمثل بالمعادلة (14):

$$out1 = LayerNormalization(\ inputs + attn\_output\ ) \quad (14)$$

وطبقة التساقط (Dropout) تُستخدم لتقليل الاختلاط العشوائي وتحسين قدرة النموذج على التعميم من خلال تطبيقا على الإخراج من الانتباه المتعدد الرؤوس والشبكة التغذوية الأمامية.تُ مثل بالمعادلة (15):

$$Output = LayerNormalization(\ out1+ffn\_output\ ) \quad (15)$$

يتم تطبيق عمليات الانتباه والكاملة المتابعة على النتائج المتسلسلة لتحسين التمثيل العميق.

تستخدم الطبقة الرابعة Conv 1D في استخراج المميزات الهامة من النصوص ولتحليل السياقات المحلية واستخلاص المعلومات الهامة من النصوص وتقوم بتطبيق الفلترة النوافذ القصيرة على البيانات المدخلة من خلال استخدام وظيفة الانتشار (Padding) والتنشيط (Activation) ووظائف التجميع (Strides).

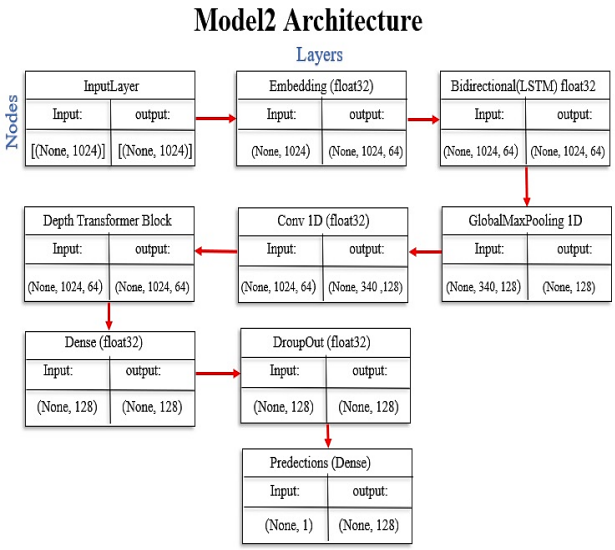
بعد ذلك تعمل الطبقة الخامسة Global Max Pooling بعد تطبيق Conv 1 على تجميع القيم القصوى في كل قناة من قنوات Conv للحصول على تمثيلات مضغوطة وشاملة للنصوص وبتقليل الأبعاد للحصول على ميزات أكثر أهمية.

ويتم تحويل الميزات المستخرجة إلى توقعات نهائية عن طريق الطبقة السادسة والمتمثلة في الطبقات المتصلة Dense والتي تحتوي على طبقات متصلة كاملة الاتصال على وحدات متصلة بالكامل (Fully Connected) مع وظائف التنشيط لتحسين القدرة التمثيلية للنموذج وتحقيق تصنيف النصوص.

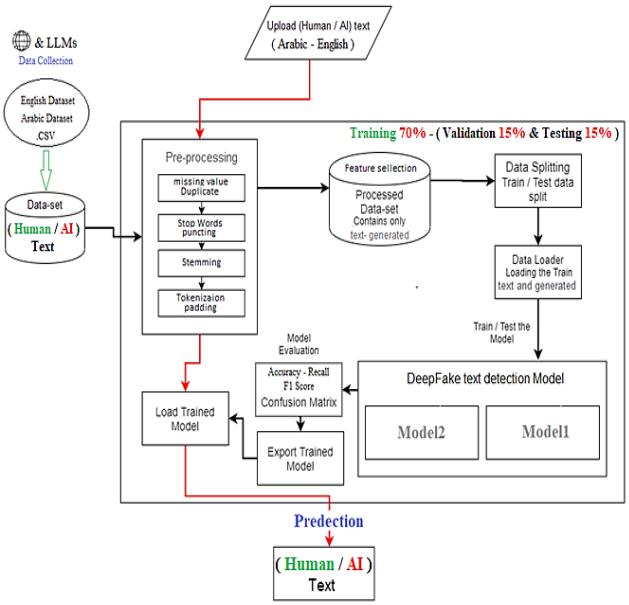
ومروراً بالطبقة السابعة Dropout والمستخدمة في لتقليل التشويش الزائف وتحسين قدرة النموذج على التعميم.

وأخيراً طبقة الإخراج Output والتي تقوم بتوليد توقعات النموذج النهائية بناءً على التمثيلات المتصلة لتقوم بإخراج التنبؤات حول ما إذا كان النص مزيفًا أم لا باستخدام وظيفة التنشيط sigmoid.

وبشكل عام، إن استخدام هذه الطبقات والتقنيات المتنوعة يمكن للنموذج تعزيز دقة كشف النصوص المزيفة العميقة باستخدام التعلم العميق وتقنيات المعالجة المسبقة.



شكل4. بنية النموذج الثاني المقترح



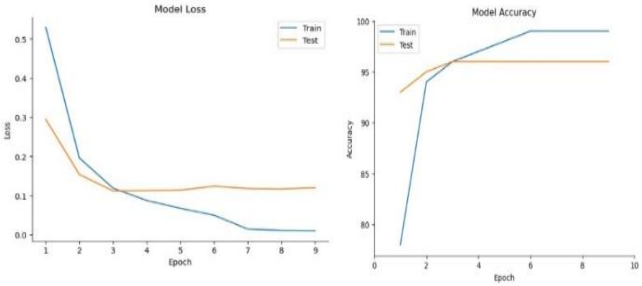
شكل5: يوضح عمل النماذج المستخدمة بشكل عام

7. النتائج والمناقشة

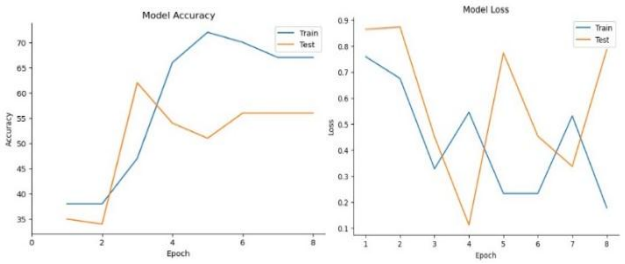
من خلال تقييمنا لأداء النموذجين السابقين على مجموعتي بيانات منفصلتين والتي تم تفصيلها سابقاً فقد أظهرت النتائج تحسناً كبيرًا وملحوظاً في دقة التصنيف والكشف وفيما يلي نستعرض بشكل مبسط نتائج التقييم لكلا النموذجين:

أ. النموذج الأول

تم تدريب النموذج الأول باستخدام مجموعة البيانات الإنجليزية ومجموعة البيانات العربية بشكل منفصل ومن خلال التقييم أظهرت النتائج دقة بلغت 96% في كشف النصوص الانجليزية المزيفة العميقة و 56% للنصوص العربية وهذا مادفعنا لاستنتاج أن هذا النموذج لم يكن أداؤه جيداً في كشف النصوص العربية وبالنظر للاشكال (6) و (7) و (8) و (9) التالية يمكن متابعة واستنتاج ما تم ذكره. ويمكن أن يرجع أحد أسباب هذا التفاوت في الأداء بين اللغتين لاختلاف التركيب اللغوي والأنماط اللغوية بينهما وخصوصاً مع النصوص الكبيرة.



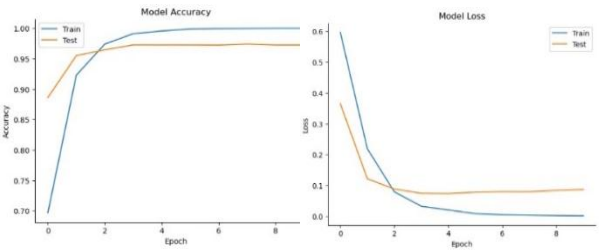
شكل7,6: يعرضان الدقة والخسار لمجموعة النصوص الإنجليزية في النموذج الأول



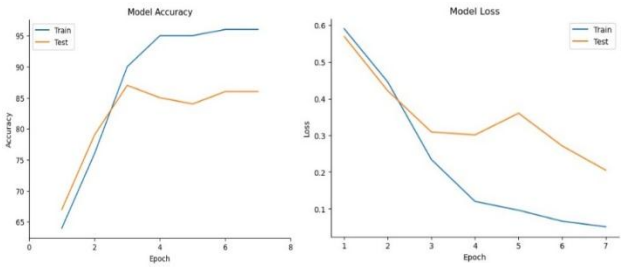
شكل9,8: يعرضان الدقة والخسار لمجموعة النصوص العربية في النموذج الأول

ب. النموذج الثاني

بما أن النموذج الأول كان أداؤه ضيفاً فيما يتعلق بمجموعة بيانات النصوص العربية بناءاً عليه قمنا ببناء وبتطوير نموذج ثاني محاولة منا الرفع من دقة كشف النصوص المزيفة العربية المزيفة العمية ولمقارنته فيما يتعلق بالنصوص الإنجليزية وبعد انتهائنا من تدريبه وتقييمه واختباره وجدنا أن النتائج كانت جيدة وافضل من النموذج الأول فقد حقق هذا النموذج دقة تبلغ حوالي 97% في كشف النصوص الانجليزية المزيفة العميقة و 86% للنصوص العربية وهذا يؤكد كفاءة هذا النموذج مقارنة بالنموذج الأول.
ومن النتائج المذكورة نستنتج أن النموذج الثاني قد حقق أداءً أفضل في كشف النصوص المزيفة مقارنةً بالنموذج الأول في كلتا المجموعتين اللغوية. على الرغم من أن النموذج الأول حقق دقة مرتفعة في الكشف عن النصوص المزيفة في اللغة الإنجليزية، إلا أن أدائه في اللغة العربية كان منخفضًا. ومع ذلك، تمكن النموذج الثاني من تحسين الدقة في كلا المجموعتين اللغوية، حيث حقق نتائج أفضل على النصوص العربية مقارنةً بالنموذج الأول.



شكل11,10: يعرضان الدقة والخسار لمجموعة النصوص الانجليزية في النموذج الثاني



شكل13,12: يعرضان الدقة والخسار لمجموعة النصوص العربية في النموذج الثاني

هذا يؤكد ان استخدام المعالجة المسبقة بشكل مناسب لكل لغة يساهم في الرفع من جودة التمثيلات المستخدمة في التدريب واستخدام تقنيات Transformer والطبقات العميقة الأخرى والتي أثبتت فعاليتها في فهم السياقات اللغوية المعقدة واستخراج العلاقات بين الكلمات بشكل أفضل.

8. الخاتمة

وختاماً، هدفت هذه الدراسة إلى تطوير نموذج قادر على كشف النصوص العربية والانجليزية المزيفة العميقة باستخدام مجموعتي بيانات تحتوي كل منهما على محتوى نصي متنوع حيث قدمنا نموذجين مختلفين نجحنا في تطبيق تقنيات المعالجة المسبقة للبيانات وبناء وتشغيل النماذج وتدريبها واختبارها على مجموعتي البيانات كلا على حدى الذي وفر رؤى ثمينة حول نقاط القوة والضعف لكل نموذج. وقد أظهرت النتائج التي تم الحصول عليها أن النموذج الثاني حقق أداءً أفضل في كشف النصوص المزيفة مقارنةً بالنموذج الأول في كلا لمجموعتي البيانات حيث تم تحسين أداء النموذج الثاني باستخدام التقنيات المسبقة وتعديل وضبط واضافة طبقات للنماذج. بشكل عام، سلطت هذه الدراسة الضوء على إمكانات استخدام تقنيات معالجة البيانات المسبقة جنباً إلى جنب مع نماذج التعلم العميق وتقنيات المحولات في كشف النصوص المزيفة العميقة.

المراجع

[1] Y. Li, Q. Li, L. Cui, W. Bi, L. Wang, L. Yang, S. Shi, Y. Zhang. 2023. "Deepfake Text Detection in the Wild," Zhejiang University, Westlake University, The University of Hong Kong, Tencent AI Lab, arXiv:2306.03643.
[2] J. Pu, Z. Sarwar, S.M. Abdullah, A. Rehman, Y. Kim, P. Bhattacharya, M. Javed, B. Viswanath. 2022 "Deepfake Text Detection: Limitations and Opportunities," Virginia Tech, University of Chicago, LUMS Pakistan, University of Virginia, arXiv:2210.09421 , In 2023 IEEE Symposium on Security and Privacy (SP), pages 19–36. IEEE Computer Society.
[3] Y. Li, Y. Chen, X. Wang, and Z. Wang. 2023. "Deepfake Text Detection: A Comprehensive Survey," arXiv:2305.01973.
[4] L. Cui, Y. Li, Q. Li, W. Bi, L. Wang, L. Yang, S. Shi, and Y. Zhang. 2023. "Deepfake Text Detection: A Literature Review," arXiv:2306.00123.
[5] K. Hayawi, S. Shahriar, and S.S. Mathew. 2023. "A Survey of Deepfake Text Detection Techniques," arXiv:2307.00546.
[6] M. Dhaini, W. Poelman, and E. Erdogan. 2023. "A Survey of Deepfake Text Detection Methods," arXiv:2308.00987.
[7] B. Brundage, M. Amodei, D. Amodei, C. Olah, J. Steinhardt, and P. Christiano. 2018. “The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,” arXiv:1802.07228.
[8] D. Jin, Z. Jin, J.T. Zhou, P. Szolovits. 2020. "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment," Computer Science & Artificial Intelligence Laboratory, MIT, University of Hong Kong, A*STAR, Singapore, arXiv:1907.11932v6.
[9] W. Zhong et al. 2020. "Neural Deepfake Detection with Factual Structure of Text," Sun Yat-

[20] C. C. Ki Chan, V. Kumar, S. Delaney, and M. Gochoo. 2020. ‘‘Combating deepfakes: Multi-LSTM and blockchain as proof of authenticity for digital media,’’ in Proc. IEEE/ITU Int. Conf. Artif. Intell. Good (AI4G), Sep. 2020, pp. 55–62.
[21] Roy SS, Roy A, Samui P, Gandomi M, Gandomi AH. 2023. Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach. IEEE Transactions on Computational Social Systems.
[22] Roy SS, Awad AI, Amare LA, Erkihun MT, Anas M. 2022. Multimodel phishing url detection using lstm, bidirectional lstm, and gru models. Future Internet.14(11):340.
[24] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On. 2020. ‘‘Fake news stance detection using deep learning architecture (CNNLSTM),’’ IEEE Access, vol. 8, pp. 156695–156706.
[24] Feng Wei and Uyen Trang Nguyen. 2019. Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), pages 101–109. IEEE.
[25] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. 2017. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.
[26] Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia. 2017. "Attention is All you Need" (PDF). Advances in Neural Information Processing Systems. 30. Curran Associates, Inc.
[27] Devlin J, Chang MW, Lee K, Toutanova K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805.
[28] Keya, A. J., Shajeeb, H. H., Rahman, M. S., & Mridha, M. F. 2023. FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. PLOS ONE, 15(12), e0294701.
[29] Zhao, D., Jiang, R., Feng, M., Yang, J., Wang, Y., Hou, X., & Wang, X. 2022. A deep learning algorithm based on 1D CNN-LSTM for automatic sleep staging. *Technol Health Care*, 30(2), 323–336. doi: 10.3233/THC-212847.
[30] Z. Liu, X. Kang, and F. Ren, 2022. “Improving speech emotion recognition by fusing pre-trained and acoustic features using transformer and BiLSTM,” in Intelligent Information Processing XI, pp. 348–357, Springer, Qingdao, China.
sen University, Microsoft Research, Fudan University, arXiv:2010.07475v1.
[10] D. Guera and E. J. Delp, “Deepfake video detection using recurrent ¨ neural networks,” in 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp. 1–6.
[11] G. Van Houdt, C. Mosquera, and G. Nápoles. 2020. "A Review on the Long Short-Term Memory Model." Artificial Intelligence Review, vol. 53, no. 1, December 2020, DOI: 10.1007/s10462-020-09838-1.
[12] A. Graves. 2012. “Long Short-Term Memory,” Berlin, Heidelberg, Springer, pp. 37–45. DOI: doi.org/10.1007/978-3-642-24797-2_4.
[13] Y. Dou, M. Forbes, R. Koncel-Kedziorski, N. A. Smith, and Y. Choi. 2022. "Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text," In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7250–7274.
[14] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. CoRR, abs/2301.11305.
[15] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. Roft: A Tool for Evaluating Human Detection of Machine-Generated Text. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations, pages 189–196.
[16] Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, and Chris Callison-Burch. 2023. Real or Fake Text?: Investigating Human Ability to Detect Boundaries between Human-Written and Machine Generated Text. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):12763–12771.
[17] K. Hayawi, S. Shahriar, S.S. Mathew. 2023. "The Imitation Game: Detecting Human and AI-Generated Texts in the Era of Large Language Models," College of Interdisciplinary Studies, Computational Systems, Zayed University, Abu Dhabi, UAE.
[18] Kadhim, A. I. 2018. An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 6, June. Department of Computer Science, College of Medicine, University of Baghdad, Iraq.
[19] S. Hochreiter and J. Schmidhuber. 1997. “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780.

[31] Zhinan Gou and Yan Li. 2023. "Integrating BERT Embeddings and BiLSTM for Emotion Analysis of Dialogue." Computational Intelligence and Neuroscience, vol. 2023, Article ID 6618452, 8 pages.
[32] Keya, A. J., Shajeeb, H. H., Rahman, M. S., & Mridha, M. F. 2023. FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. Technol Health Care, 30(2), 323–336. doi: 10.1371/journal.pone.029470.
[33] Kaliyar RK, Goswami A, Narang P, Sinha S. 2020, FNDNet–a deep convolutional neural network for fake news detection. Cognitive Systems Research. 2020;61:32–44.
[34] J. Pennington, R. Socher, and C. D. 2014. Manning, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.