تقنيات معالجة اللغة الطبيعية لأغراض البحث والاسترجاع في مجال المكتبات والمعلومات*

إعداد

د. مصطفى محمد إبراهيم الهلالي

قسم المكتبات والوثائق والمعلومات

كلية الآداب- جامعة القاهرة

mostafaelhelalyy@cu.edu.eg

مراجعة وإشراف

أ.د. أسامة أحمد جمال القلش

أستاذ المكتبات والمعلومات

كلية الآداب - جامعة القاهرة

alqlsh@yahoo.com

المستخلص:

تعد تقنية معالجة اللغة الطبيعية فرع من تقنيات الذكاء الاصطناعي، والتي بدورها جعل التعامل مع الحاسوب يكون بلغة أقرب إلى اللغة الطبيعية، حيث سعت الدراسة إلى التعريف بتقنية معالجة اللغة الطبيعية، وعرض تاريخها منذ الستينيات حتى الآن، وتوضيح المصطلحات الأساسية المستخدمة في مجال معالجة اللغة الطبيعية، فضلًا عن تحديد كل من: العناصر التكوينية لتقنية معالجة اللغة الطبيعية، والمستويات اللغوية في مجال معالجة اللغة الطبيعية، والمراحل التي تمر بها تقنية معالجة اللغة الطبيعية، والوصول إلى تطبيقات معالجة اللغة الطبيعية في علم المكتبات والمعلومات. واعتمدت الدراسة على المنهج الوصفي التحليلي في

* بحث مقدم ضمن متطلبات الحصول على درجة الدكتوراه لرسالة بعنوان: " دور الأنطولوجيات في تمثيل المعرفة بتقنيات الذكاء الاصطناعي لأغراض البحث والاسترجاع: دراسة تطبيقية على الإنتاج الفكري في مجال التربية"؛ إشراف أ. د. أسامة أحمد جمال القلش- قسم المكتبات والوثائق والمعلومات، كلية الآداب، جامعة القاهرة.

تاريخ استلام البحث: 2024/05/14 تاريخ إجازة البحث: 2024/09/20

مراجعة الإنتاج الفكري الراجع والجاري لحصر المفاهيم المتعلقة بمجال معالجة اللغة الطبيعية بالاعتماد على قواعد البيانات الأجنبية المتاحة على بنك المعرفة المصري Egyptian Knowledge Bank (EKB)، وقد توصلت الدراسة إلى بعض النتائج، أبرزها: التوقع بظهور الكثير من البرمجيات ذات الواجهات الرسومية Graphic User Interface، التي ستسمح باستخدام تقنيات معالجة اللغة الطبيعية دون الحاجة إلى استخدام أكواد Encoding، مما يساعد المستخدمين المبتدئين في تطبيق تقنيات معالجة اللغة الطبيعية بسهولة.

الكلمات المفتاحية: معالجة اللغة الطبيعية، فهم اللغة الطبيعية، الذكاء الاصطناعي، استرجاع المعلومات، استرجاع النصوص.

أولًا: الإطار المنهجي للدراسة:

1/1. التمهيد.

بدأت تقنية معالجة اللغات الطبيعية باعتبارها فرعًا من فروع الذكاء الاصطناعي، حيث يكمن هدفها في جعل التعامل مع الحاسوب بلغة أقرب إلى اللغة الطبيعية، والاستفادة منها لهدف معلوم ومحدد مسبقًا، ويرجع ذلك إلى أن فهم النص حاسوبيًا - أي التمثيل النطقي الكامل للنص- بكل مراحل المعالجة للنص التربوي تعد ضربًا من المستحيل أو شبه المستحيل، فلا وجود لنظام متكامل لمعالجة اللغة العربية يشمل كل مراحل المعالجة حتى الآن، لا سيما أننا ما زلنا نتحدث عن المعالجة وليس الفهم، ومع ذلك فلا وجود لنظام يحتوي على كل مراحل المعالجة؛ لأننا ببساطة شديدة لا نعرف كيف تتم عملية الفهم حتى الآن، حيث إن هناك العديد من النظريات من أجل ذلك، ومهمة الحاسوب تتمثل في برمجة هذه النظريات، والتأكد من صحتها (Chopra et al.، 2013).

شكل (1) معالجة اللغة الطبيعية

يوضح الشكل السابق أن هناك:

1- **فهم للغة الطبيعية Natural Language Understanding:** مهمتها هي الفهم، حيث تكون المدخلات عبارة عن مصطلحات طبيعية في مجال التربية (مجال الدراسة الموضوعي).

2- **توليد للغة الطبيعية Natural Language Generation:** يعد التوليد جيل فرعي من معالجة اللغة الطبيعية، ويشار إليه أيضًا بمصطلح "توليد النص Text Generation" (**Chopra et al.، 2013**).

فالهدف الرئيسي من تقنية معالجة اللغة الطبيعية؛ هو تحقيق معالجة للغة تشبه إلى حد كبير معالجة اللغة البشرية لعدة مهام أو تطبيقات، فضلًا عن إمكانية تحليل النصوص من خلال تطبيقات متعددة سيتم تناولها خلال هذه الدراسة.

2/1. مصطلحات الدراسة:

1/2/1. معالجة اللغة الطبيعية Natural Language Processing:

عرفت عفاف سفر السلمي (2017) معالجة اللغة الطبيعية في دراستها المعنونة (تطبيقات الذكاء الاصطناعي لاسترجاع المعلومات في جوجل) بأنها: "التفاعل البشري بين الذكاء الاصطناعي والحاسبات عن طريق المعالجة الطبيعية للغة؛ للجمع بين التعلم الإنساني ومنطق الآلة، فضلًا عن خلق برامج الكمبيوتر، التي توفر التفاعل بين الإنسان والحاسب الآلي لتخزين المعلومات الأولية وحل مشاكل محددة، والقيام بالمهام المتكررة التي يطلبها المستخدم، والقيام بعدد من الوظائف، مثل: تصحيح الأخطاء الإملائية، تشكيل الهيكل النحوي للجمل، توفير علاقة دلالية.

2/2/1. الذكاء الاصطناعي Artificial Intelligence:

قدم قاموس جمعية المكتبات الأمريكية *ALA Glossary of library & Information Science* في طبعته الإلكترونية المتاحة على الويب الصادرة في (9 مايو 2021) تعريفًا للذكاء الاصطناعي بأنه: "عبارة عن الأجهزة والتطبيقات المصممة لتكرار القدرات البشرية عن كثب، فضلًا عن استخدام إمكانية استخدام تقنيات الذكاء الاصطناعي في العديد من التقنيات، مثل: التعرف على الصوت *Voice Recognition*، والأنظمة الخبيرة *Expert Systems*، ومعالجة اللغة الطبيعية *natural language processing*.

3/2/1. تمثيل المعرفة Knowledge Representation:

يعد تمثيل المعرفة مجال من مجالات الذكاء الاصطناعي؛ حيث يهتم بتقديم وعرض المعلومات في شكل يمكن للحاسوب فهمه واستخدامه لحل المشكلات ومعالجة المهام المختلفة، كما يقدم تمثيل المعرفة للآلات *Machines* القدرة على التفكير والتصرف مثل البشر في عمليات الفهم والتفسير والاستدلال، ويتم ذلك من خلال تدريب آلة الذكاء الاصطناعي على التعلم من المعلومات المتاحة أو الخبرة أو الخبراء (*Fingnet*، 2023).

3/1. مشكلة الدراسة ومبررات اختيارها:

نشأت مشكلة هذه الدراسة من خلال الملاحظات الآتية:

أولًا: تعد قضية إكساب نظم الاسترجاع القدرة على فهم استفسارات المستفيدين إحدى المشكلات الرئيسة في جوهر البحوث والدراسات في هذا الصدد، وذلك إلى جانب قضايا التعامل

مع التحديات الخوارزمية والاسترجاعية واللغوية، وغيرها التي تقف وراء عجز نظم الاسترجاع النصية التقليدية في الوفاء بتلبية احتياجات المستفيدين والباحثين في توفير القدر الوافي من المطابقة بين الاحتياجات المعبر عنها والنتائج المسترجعة.

ثانيًا: هناك العديد من الجهود الرامية إلى إكساب الآلة القدرة على فهم ما تتلقاه وتعالجه وتسترجعه من نصوص، وتأتي الأنطولوجيات *Ontology* على رأس هذه الجهود كأحد أبرز نظم تمثيل المعرفة *Knowledge Representation* للآلة باستخدام تقنية معالجة اللغة الطبيعية *Natural Language Processing*.

ثالثًا: ثمة جهود رامية إلى الدمج المنهجي للذكاء الاصطناعي في المكتبات ومؤسسات المعلومات باعتبار أن ذلك يعزز القدرة على **مواجهة التحديات** في هذا المجال، فضلا عن **ابتكار ممارسات** تسهم في النهاية بتسريع التقدم نحو تعزيز القدرة على الاسترجاع الفعال في أنظمة استرجاع المعلومات المتخصصة.

لذلك كان ينبغي التعرف على تقنية معالجة اللغة الطبيعية كأحد تقنيات الذكاء الاصطناعي، والتي يمكن الاستفادة منها في تطوير المكتبات ومؤسسات المعلومات بغرض إكساب نظم الاسترجاع في هذا المجال القدرة على فهم استفسارات المتخصصين، ومن ثم تحقيق نتائج استرجاعية أكثر كفاءة وفعالية.

4/1. أهمية الدراسة:

تستمد الدراسة أهميتها من الصعوبات والمشكلات التي تواجهها نظم البحث في الاسترجاع، حيث ترتبط أهمية الدراسة بأهمية تقنية معالجة اللغة الطبيعية في تطوير وتحسين أداء المكتبات ومؤسسات المعلومات من خلال توظيف واستثمار تطبيقات الذكاء الاصطناعي فيما يتعلق بمعالجة اللغة الطبيعية.

5/1. أهداف الدراسة:

تهدف الدراسة إلى التعرف على تقنية معالجة اللغة الطبيعية كأحد تقنيات الذكاء الاصطناعي، حيث تسعى الدراسة بصفة عامة إلى التعريف بمفهوم معالجة اللغة الطبيعية ودورها في تحسين أداء المكتبات ومؤسسات والمعلومات.

وينبثق عن هذا الهدف؛ عدد من الأهداف الفرعية والتي تتمثل في:

1- التعريف بتقنية معالجة اللغة الطبيعية وتاريخها.

2- تحديد المصطلحات الأساسية المستخدمة في مجال معالجة اللغة الطبيعية.
3- التعرف على العناصر التكوينية لتقنية معالجة اللغة الطبيعية.
4- تحديد المراحل الأساسية لتطبيق تقنية معالجة اللغة الطبيعية.
5- التعرف على الخوارزميات المستخدمة عند تطبيق تقنية معالجة اللغة الطبيعية.
6- الإفادة من تقنيات معالجة اللغة الطبيعية في المكتبات ومؤسسات المعلومات.

6/1. تساؤلات الدراسة:

1- ما المقصود بتقنية معالجة اللغة الطبيعية؟ وما تاريخها؟
2- ما المصطلحات الأساسية المستخدمة في مجال معالجة اللغة الطبيعية؟
3- ما آليات تطبيق تقنية معالجة اللغة الطبيعية؟
4- ما الخوارزميات المستخدمة عند تطبيق تقنية معالجة اللغة الطبيعية؟
5- كيف يمكن الإفادة من تطبيقات تقنيات معالجة اللغة الطبيعية في المكتبات ومؤسسات المعلومات؟

7/1. مجال الدراسة وحدودها:

الحدود الموضوعية:

تركز الدراسة على تقنية معالجة اللغة الطبيعية كأحد تقنيات الذكاء الاصطناعي لما لها من انعكاسً في رفع كفاءة وفاعلية الاسترجاع في المكتبات ومؤسسات المعلومات.

الحدود النوعية:

تتمثل الحدود النوعية في التعرف على تقنية معالجة اللغة الطبيعية بالاعتماد مختلف أنواع مصادر المعلومات من كتب، ومقالات، وأعمال مؤتمرات، ورسائل جامعية.

الحدود اللغوية:

تركز الحدود اللغوية للدراسة على التعامل مع الإنتاج الفكري الصادر باللغة الإنجليزية حول ظاهرة الدراسة، وما صدر بخلاف اللغة الإنجليزية، تقوم الدراسة بالتعامل معه على صعيد مستخلصه العلمي.

الحدود الزمنية:

تتمثل في رصد محاولات التأصيل النظري لتقنية معالجة اللغة الطبيعية، منذ صدور أول عمل علمي محكم تناول هذه الظاهرة عام 2003 للباحث **Chowdhury** في دراسة علمية له تحت عنوان "معالجة اللغة الطبيعية Natural Language Processing "، حتى وقت الانتهاء من الدراسة.

8/1. منهج الدراسة وأدواتها:

تعتمد الدراسة على **المنهج الوصفي التحليلي**، وذلك لمراجعة الإنتاج الفكري الراجع والجاري لحصر المفاهيم المتعلقة بتقنية معالجة اللغة الطبيعية، بالاعتماد على قواعد البيانات الأجنبية المتاحة على بنك المعرفة المصري *Egyptian Knowledge Bank (EKB)*، فضلًا عن الاعتماد على بعض الأدوات الأخرى، مثل: (الباحث العلمي *Google Scholar*، ومحرك بحث جوجل *Google*، وبوابة الأبحاث *Research Gate*).

ثانيًا: الإطار النظري للدراسة:

1/2. تعريف تقنية معالجة اللغة الطبيعية.

قدم قاموس أكسفورد Oxford Dictionary تعريفًا لتقنية معالجة اللغة الطبيعية بأنها: "عملية فهم كيفية استخدام النصوص والمواد المماثلة من قبل الأنظمة المحوسبة وكيفية تشغيلها على الحاسبات الآلية"، حيث يمكن تطبيق تقنيات معالجة اللغة الطبيعية في تحليل اللغة الطبيعية والكلام الوارد داخل النصوص المختلفة.

وقد طُبقت أساليب معالجة اللغة الطبيعية في البداية على اللغات المهددة بالانقراض؛ لمنع انقراضها، بيد أنها طبقت واستخدمت -هذه الأساليب- مؤخرًا في العديد من الدراسات؛ لتنظيم وفهم البيانات الكبيرة، حاليًا سيكون من الأصعب من حيث الوقت والأكثر تكلفة العمل دون تقنية معالجة اللغة الطبيعية في العديد من المجالات، بما في ذلك تلك المجالات (التسويق، التحقق من المعلومات، استرجاع المعلومات، ...إلخ) (Taskin & Al، 2019).

يعد مجال معالجة اللغة الطبيعية تخصصًا يركز على فهم وتوليد اللغة الطبيعية من قبل الآلات باستخدام الخوارزميات المختلفة، لذلك يمكننا القول بإن تقنية معالجة اللغة الطبيعية حقًا في الواجهة بين علوم الكمبيوتر computer science، وعلوم اللغة linguistics؛ نظرا لقدرة تقنية معالجة اللغة الطبيعية على التفاعل المباشر بين الآلة والبشر، ودورها في

مساعدة الآلة في إدراك معاني الكلمات، وكيانات المصطلحات الواردة داخل النصوص (Daniel، 2023).

2/2. تاريخ تقنية معالجة اللغة الطبيعية.

هناك عدة مراحل رئيسية مرت بها تقنية معالجة اللغة الطبيعية في تاريخ معالجة اللغة الطبيعية، وهي:

1/2/2 المرحلة الأولى: ترجمة الآلة (قبل الستينيات).

قُدِمّ مفهوم معالجة اللغة الطبيعية في القرن السابع عشر من قِبَلِ الفيلسوف والرياضياتي " جو تفريد فيلهلم لايبنز" Gottfried Wilhelm Leibniz (1646-1716) وعالم الرياضيات رينيه ديكارت Descartes René (1596-1650) في دراساتهما بشأن العلاقات بين الكلمات واللغات، التي شكلت أساسًا لتطوير محرك ترجمة اللغات (Santilal، 2020).

فضلًا عن تقديم أول براءة اختراع مرتبط بترجمة الآلة من قبل المخترع والمهندس جورج أرتسروني Georges Artsrouni في عام 1933، لكن الدراسة الرسمية والبحث قدما من قبل آلان تورينج Alan Turing في مقالته الصادرة بعنوان "آلات الحوسبة والذكاء Computing Machinery and Intelligence" عام 1950، حيث تناول المقال اختبار تورينج الشهير الذي استخدم رسميًا كمعيار تقييم للذكاء الآلي، منذ أن كان البحث والتطوير في معالجة اللغة الطبيعية مركزًا بشكلٍ رئيسٍ على ترجمة اللغات في ذلك الوقت (Turing، 1950).

عُقِدَ المؤتمر الدولي الأول والثاني حول ترجمة الآلة في عامي 1952 و1956 باستخدام تقنيات أساسية قائمة على القواعد basic rule-based Techniques. حيث شاركت تجربة جورج تاون- أي بي إم Georgetown-IBM عام 1954 في ترجمة الآلة الكاملة تلقائيًا لأكثر من 60 جملةً روسية إلى الإنجليزية، مما دعا إلى التفاؤل بشكلٍ زائدٍ بأنه يمكن حل مشكلة ترجمة الآلة بالكامل خلال بضع سنوات.

ومع ذلك، تم تحقيق اختراق في معالجة اللغة الطبيعية من قِبَلِ البروفيسور المتقاعد نوام تشومسكي Noam Chomsky في عام 1957، لكن منذ نشر تقرير اللجنة الاستشارية لمعاجلة اللغات التلقائية "ALPAC " في عام 1966، فقد كَشَفَ عن عدمِ تقدمٍ كافٍ في مجال الذكاء الاصطناعي وترجمة الآلة (Lee، 2023).

2/2/2 المرحلة الثانية: الذكاء الاصـطناعي المبكر في مجال معالجة اللغة الطبيعية (1960م – 1970م).

كان التطوير الرئيس لمجال معالجة اللغة الطبيعية يركز على كيفية اسـتخدامه في مجالات مختلفة مثل هندسـة المعرفة Knowledge Engineering، وبعد انتشـار الذكاء الاصـطناعي بشـكلٍ شـائعٍ مع مرور الوقت، ظهر نظام البيسـبول Baseball System كمثالٍ نموذجي على نظام خبير في مجال الأسـئلة والأجوبة، للتفاعل بين الإنسـان والكمبيوتر، الذي طُور في السـتينيات، ولكن كانت الإدخالات مقيدة، وظلت تقنيات معالجة اللغة الطبيعية في مستواها البسيط (Green et al.، 1961).

في عام 1968، طور البروفيسـور مارفن مينسـكي Marvin Minsky (1927-2016) نظامَ معالجةٍ لغويةٍ طبيعيةٍ أكثرَ قوةٍ، حيث اسـتخدم نظامًا متقدمًا يعتمد على الذكاء الاصـطناعي في الربط بين البشـر والكمبيوتر؛ لتوفير تفسـيرات مبنية على المعرفة للأسـئلة والأجوبة التي يتلقاها النظام، فضـلًا عن ذلك اقترح البروفيسـور ويليام أ. وودز William A. Woods شـبكةَ ترجمةٍ معززةٍ augmented translation network (ATN)؛ لتمثيل الإدخال باللغة الطبيعية في عام 1970.

خلال هذه الفترة، بدأ العديد من المبرمجين في تحويل الأكواد إلى لغات ذكاء اصـطناعي مختلفة؛ لتصـور معرفة المفاهيم الخاصـة باللغة الطبيعية حول المعلومات الهيكلية الحقيقية في العالم، إلى وضع فهم بشري تستطيع الآلة استيعابه، ومع ذلك لم تتمكن هذه النظم الخبيرة من تلبية التوقعات (Lee، 2024).

3/2/2 المرحلـة الثـالثـة: المنطق النحوي Grammatical Logic في مجـال معـالجـة اللغـة الطبيعية (1970م – 1980م).

بدأ في هذه المرحلة ظهور مصـطلح تمثيل المعرفة Knowledge Representation، حيث تحوّل البحث هنا إلى تمثيل المعرفة، ومنطق البرمجة، والاسـتدلال في مجال الذكاء الاصـطناعي (Hausser، 2014).

كانت هذه الفترة تُعَدُّ مرحلة المنطق النحوي في مجال معالجة اللغة الطبيعية في التعامل مع الجمل والعبارات المختلفة، حيث استخدمت تقنيات قوية لمعالجة الجمل، مثل نظرية تمثيل الحوار discourse representation theory، تعتمد نظرية تمثيل الحوار على تفسير الجمل والعبارات، بالاعتماد على موارد وأدوات عملية مثل روبوتات محادثة للأسئلة والأجوبة Q&A chatbots (Hausser، 2014).

على الرغم من تعثر البحث والتطوير بسبب قوة مجال الحاسب الآلي في هذا الفترة، إلا أن الهدف الأساسي كان توسيع مجال معالجة اللغة الطبيعية في الثمانينات من القرن الماضي (Lee، 2024).

4/2/2 المرحلة الرابعة: الذكاء الاصطناعي وتعلم الآلة (1980م - 2000م).

في هذه المرحلة نجح نموذج شبكة هوبفيلد Hopfield Network الذي اقترحه البروفيسور المتقاعد جون هوبفيلد John Hopfield في مجال تعلم الآلة Machine Learning، حيث نشط عصرًا جديدًا من البحوث في مجال معالجة اللغة الطبيعية باستخدام تقنيات تعلم الآلة كبديل للأساليب المعقدة المعتمدة على القواعد complex rule-based Methods في العقود السابقة (Bender et al.، 2013).

كما أن التحديثات التي طرأت على تكنولوجيا الحاسب في الطاقة الحسابية والذاكرة، تكمل نظرية اللغويات لشومسكي Chomsky's theory، التي كان لها دورٌ في تعزيز معالجة اللغة الطبيعية من خلال تقنيات تعلم الآلة في علم اللغة النصية corpus linguistics، حيث تُعرف هذا المرحلة من التطور باسم "اللغويات النصية لتعلم الآلة".

وتجدر الإشارة إلى أن أهم ما يميز هذه المرحلة، ظهور مشروع IBM DeepQA الذي قاده الدكتور ديفيد فيروتشي David Ferrucci لنظامه الخاص بالإجابة على الأسئلة والاستفسارات، الذي طُور في عام 2006 (Bender et al.، 2013).

5/2/2 المرحلة الخامسة: الذكاء الاصطناعي، البيانات الضخمة، والشبكات العميقة (2010م – حتى الآن).

تتسم هذه المرحلة بالتطورات التكنولوجية الحديثة، التي طرأت على مجال الذكاء الاصطناعي وعلوم البيانات، فقد تطور مجال معالجة اللغة الطبيعية، وأصبح متضامنًا مع

تكنولوجيا الحوسبة السحابية Cloud Computing، والحوسبة المحمولة Mobile Computing، والبيانات الضخمة Big Data فيما يتعلق بتحليل الشبكات العميقة Deep Network، وقد ساهمت كل من (جوجل Google، وفيسبوك Facebook، وأمازون Amazon) في تطوير الشبكات العصبية العميقة deep neural networks في عام 2010؛ لتصميم منتجات مثل القيادة الآلية Auto-Driving، وروبوتات الدردشة للأسئلة والأجوبة Q&A chatbots، وتطوير التخزين storage development (Lee، 2024).

3/2. المصطلحات الأساسية في مجال معالجة اللغة الطبيعية Basic NLP Terminology.

هناك مجموعة من المصطلحات الشائعة والمستخدمة باستمرار في مجال معالجة اللغة الطبيعية كما ذكرها (Eisenstein، 2019) وهي:

1- **الجملة Sentence:** وحدة من وحدات اللغة الكتابية، أي الكيان الأساسي في المحادثة أو الكلام.
2- **الكلمة Utterance:** وحدة من وحدات اللغة الشفوية، حيث تختلف عن مفهوم الجملة؛ لأن عادة ما يكون الكلام موضحًا للمجال والثقافة، مما يعني أنه قد يتغير وفقًا للدول أو حتى داخل الدولة.
3- **الترميز Tokenization:** هي الكيانات العامة التي تأتي داخل النص، كما أنها تختلف عن شكل الكلمة؛ حيث يمكن أن تكون الرموز كلمات ذات معنى، أو رموز، أو علامات ترقيم، أو حروف بسيطة ومتميزة.
4- **التجذير Stemming:** يقصد بها عملية إرجاع الكلمة الواردة داخل النص إلى جذرها الأصلي، أي الشكل الأصلي للكلمة.
5- **التصريف Lemmatization:** يقصد بها عملية تجميع أشكال الكلمات المختلفة الواردة داخل النص، بحيث يمكن تحليلها كعنصر واحد، يمكن التعرف عليه من خلال تصريف الكلمة.

4/2. العناصر التكوينية لتقنية معالجة اللغة الطبيعية.

تتكون تقنية معالجة اللغة الطبيعية من ثلاثة عناصر رئيسة، هما:

1/4/2 فهم اللغة الطبيعية Natural Language Understanding (NLU):

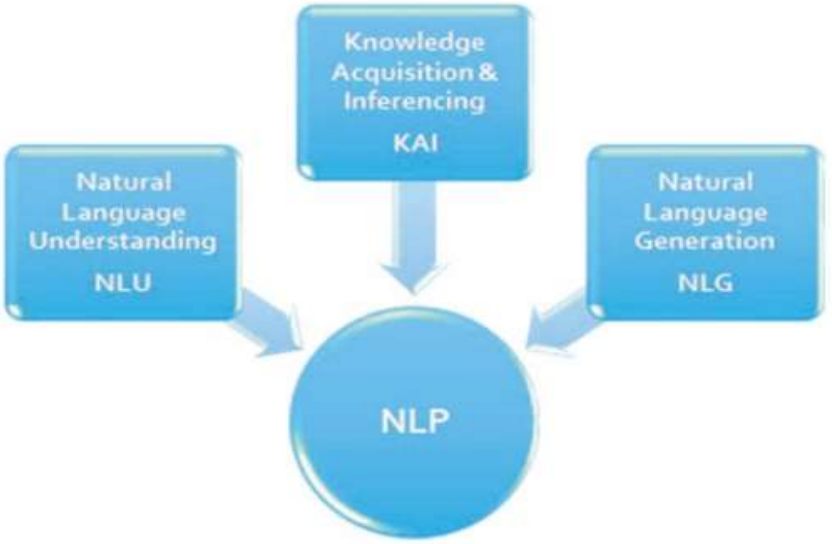
يقصد بفهم اللغة الطبيعية التقنية أو الأسلوب المصمم لفهم معاني اللغات المنطوقة للبشر من خلال تحليل البنية النحوية والدلالية للجمل والعبارات (Lee، 2023).

2/4/2 توليد اللغة الطبيعية Natural Language Generation (NLG):

يشـمل توليد اللغة الطبيعية إنشـاء إجابات Answers، واسـتجابات Responses، وتغذية مرتدة Feedback في حوار الإنسـان مع الآلة، فعملية توليد اللغة الطبيعية بمثابة عملية ترجمة آلية متعددة الجوانب، حيث تحول الاسـتجابات إلى نصـوص وجمل، عن طريق النص إلى كلام من اللغة المستهدفة، وإنتاج استجابات كلامية تقريبية للإنسان (Lee، 2023).

3/4/2 اكتساب المعرفة والاستدلال Knowledge Acquisition and Inferencing (KAI):

أما نظام اكتسـاب المعرفة والاسـتدلال؛ فيعتبر نظامًا لإنشـاء اسـتجابات مناسـبة، بعد أن تم التعرف الكامل على اللغات المنطوقة بواسـطة تقنية فهم اللغة الطبيعية NLU، إلا أن هناك مشـكلة لم تُحل بعد في اكتسـاب المعرفة والاسـتدلال فيما يتعلق بتعلم الآلة والذكاء الاصـطناعي، بواسـطة النظام القائم على القواعد التقليدية؛ نظرًا لتعقيدات اللغة الطبيعية والحوار والمحادثات (Lee، 2023).



- شكل (2) العناصر التكوينية لتقنية معالجة اللغة الطبيعية (Lee، 2023).

5/2. المستويات اللغوية في مجال معالجة اللغة الطبيعية.

تُعد المستويات اللغوية تحليلًا وظيفيًا للغات البشرية المكتوبة والمنطوقة، وبناء على ذلك هناك مستويات ستة في تحليل اللغويات في مجال اللغة الطبيعية، هما:

(الصوتيات Phonetics، علم الصوتيات Phonology، الصرف Morphology، النحو Syntax، المعاني Semantics، البرجماتية (علم الدلالة أو الاستدلال) (pragmatics (Discourse)

يعد **الهيكل اللغوي** الأساسي للغة المنطوقة، يشمل **الصوتيات وعلم الصوتيات**، حيث يفرق مصطلح الصوتيات عن علم الصوتيات في تركيز مصطلح **الصوتيات Phonetics** على دراسة نظام تصنيف الأصوات الناتجة عن الكلام المنطوق، أما مصطلح **علم الصوتيات Phonology** يهتم بدراسة علم الصوتيات بما يشمله من التاريخ ونظريات التغيرات الصوتية في اللغة أو في عدة لغات (Elsherif، 2024).

أما **هيكل اللغة** المباشر يرتبط بمستويات **الصرف والنحو**، **فالصرف** Morphology هو الشكل ومستوى الكلمة المحددة بواسطة القواعد النحوية عمومًا، حيث يشير إلى أصغر شكل في تحليل اللغويات، والذي يتكون من أصوات؛ لدمج الكلمات بوظيفة نحوية أو معجمية، أما النحو Syntax، فهو العلم الذي تعرف به الضوابط التي تحكم التراكيب اللغوية، ويترتب عليها صحة الكلام (Lee، 2023).

تتعامل **البنية المتقدمة** مع معنى اللغة الفعلي على المستويين **الدلالي والبراجماتي**، **فالمستوى الدلالي** Semantics هو مجال المعنى، الذي يتألف من الصرف والنحو، ولكن يتطلب المستوى الدلالي تخصيص المعنى الصحيح على الفور بالمفردات، وشكل المصطلح والقواعد النحوية والجملة.

ويقصد بالمستوى البراجماتي pragmatics استخدام اللغة في سياقات محددة، فلا يجب أن يكون معنى الكلمة هو نفس الشكل المجرد في الاستخدام الفعلي؛ لأنه يعتمد إلى حدٍ كبيرٍ على مفهوم الأفعال اللفظية ومحتويات البيان مع تحليل النية والتأثير في أداء اللغة، حيث يركز المستوى البراجماتي على توضيح المعنى المخفي أو غير المباشر وراء الإشارة والتعبيرات اللغوية في اللغة (Elsherif، 2024).

يوضح الشكل (3) المستويات الستة للغة الطبيعية:

شكل (3) المستويات الستة للغة الطبيعية (Lee، 2023)

6/2. مراحل تقنية معالجة اللغة الطبيعية.

تمر تقنية معالجة اللغة الطبيعية بعدة مراحل، هما:

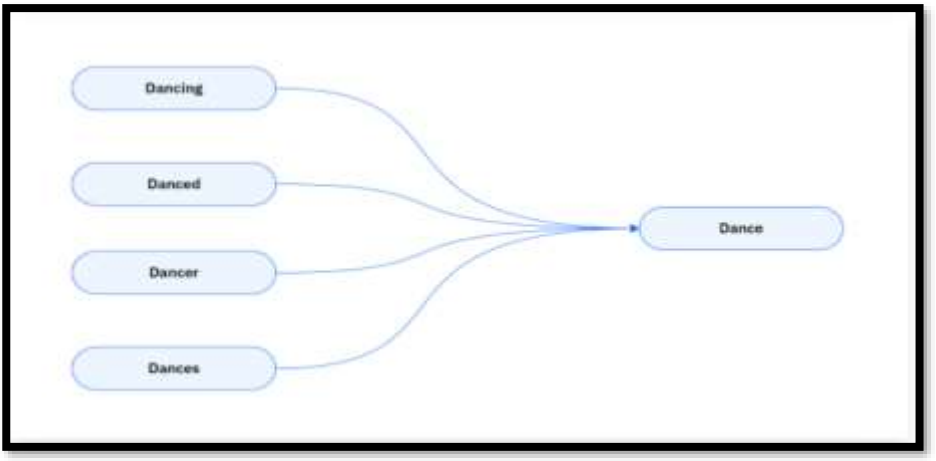
- **المرحلة الأولى: المعالجة المسبقة للنص Text Preprocessing:**

تعد المعالجة المسبقة للنص الخطوة الأولى في خط الإنتاج لمعالجة اللغة الطبيعية، والتي تتضمن سلسلة من الخطوات المصممة لتحويل النص الخام إلى تنسيق أكثر سهولة وتوحيدًا، مما يمكن الخوارزميات من معالجته؛ نظرًا للتنوع الهائل في اللغة البشرية، بما في ذلك التغييرات في البنية النحوية والقواعد اللغوية، فإن تهيئة النص عملية مهمة لتقليل التعقيد وتعزيز الكفاءة الحسابية (Brank et al.، 2016).

تعتمد المعالجة المسبقة للنص على مجموعة من التقنيات بغرض تعزيز أداء نموذج معالجة اللغة الطبيعية، مثل إزالة الكلمات غير الدالة الزائدة Stop words Removal، والتجذير Stemming، والتصريف Lemmatiozation، والتطبيع Normalization، وستتناول الخطوات القادمة كيفية إزالة الكلمات غير الدالة الزائدة، والتطبيع، أما التجذير والتصريف فهما أهم ما يميز هذه الخطوة في إعداد المعالجة المسبقة للنص (Brank et al.، 2016).

فالتجذير يعني إرجاع الكلمة لجذرها الأساسي مما يساعد في توحيد الاختلافات في الكلمة في تمثيل واحد، وتختلف خوارزميات التجذير، على الرغم من أنها تشترك في بعض وسائط العمل العامة، حيث تزيل الخوارزميات لواحق الكلمات، ثم تزيل أي سلاسل حروف ملحقة بالكلمة.

وتعد خوارزمية بورتر للتجذير Porter stemming algorithm، أكثر خوارزميات التجذير المستخدمة على نطاق واسع، والنسخة المحدثة منها هي خوارزمية السنوبول Snowball stemmer، التي تساعد في فهم تجذير الكلمات بشكل أفضل، أما التصريف يعني إرجاع الكلمة للصرف المصدري الخاص بها، حيث يسعى التصريف إلى شكل قاعدة واحدة للكلمة داخل القاموس، مما يضمن أن الكلمة الواردة داخل النص هي شكل موجود من الكلمة (Mural & Kavlakoglu، 2023).



شكل (4) التجذير والتصريف في معالجة اللغة الطبيعية (Mural & Kavlakoglu، 2023)

وترجع أهمية هذه الخطوة إلى أن البيانات الخام غالبًا ما تأتي محملة بعلامات الترقيم والرموز الخاصة، والتي قد لا تكون مفيدة لمهام معالجة اللغة الطبيعية المحددة، ففي هذه الخطوة يُقلص النص إلى محتواه الأساسي، فضلًا عن إزالة الأحرف غير المهمة، والرموز، والتنسيق من النص، مما يقلل من التشتت والتركيز على تحليل البيانات ذات الدلالة فقط (Brank et al.، 2016).

حيث تهدف المرحلة الأولى إلى إزالة المعلومات غير ذات الصلة أو الزائدة من بيانات النص، فهذه هي الخطوة الأولى الحاسمة المسؤولة عن التخلص من الضوضاء، التي قد تنحرف عن أداء نماذج معالجة اللغة الطبيعية، حيث تشكل تلك المرحلة الأساس الصلب لمعالجة اللغة الطبيعية، مما يتيح للحاسبات القدرة على فهم وتفسير لغة البشر (Mural & Kavlakoglu، 2023).

- **المرحلة الثانية: التأخيذ Tokenization:**

تعد هذه الخطوة أساسية لمعظم المهام في مجال معالجة اللغة الطبيعية، كما أنها من أهم الخطوات الأولى عند بناء نموذج لمعالجة اللغة الطبيعية؛ لأنها تؤثر مباشرة على قدرة النموذج على التعلم من بيانات النص، حيث تحول النص الخام إلى تنسيق يمكن للخوارزميات التعامل معه بشكلٍ أكثرَ فعالية، فهي عملية تقسيم النص إلى وحدات أصغر وعناصر ذات معنى (Brank et al.، 2016).

تعتمد هذه الخطوة على التعامل مع استخدام المسافات البيضاء، والترقيم، والقواعد اللغوية داخل النص؛ لتحديد حدود الكلمات بدقة، وهذا ينطوي على تحديد حدود الجمل، والتي يمكن أن تكون تحديًا بسبب استخدام الترقيم لأغراض أخرى غير تحديد الجمل (Mural & Kavlakoglu، 2023).

تعتمد كفاءة وفعالية تلك المرحلة بشكلٍ كبيرٍ على مدى نجاح عمليات تنقيح النص Text Cleaning، ويقصد بها إزالة الأحرف غير المهمة والرموز والتنسيقات داخل النص، مما يضمن أن تركز فقط عملية التأخيذ على العناصر ذات الدلالة في النص، حيث تركز عملية التأخيذ على استكشاف أساليب لفهم كيفية تقسيم النص بشكلٍ فعالٍ إلى مكونات قابلة

للتحليل، واختيار الأسلوب يعتمد على المتطلبات المحددة للمهام المطلوبة في مجال معالجة اللغة الطبيعية (Elsherif، 2024).

- **المرحلة الثالثة: إزالة الكلمات غير الدالة Stop Word Removal:**

إن الكلمات غير الدالة عبارة عن مجموعة من الكلمات المستخدمة بشكلٍ شائعٍ في اللغة، والتي يمكن استبعادها دون التأثير على محتوى ومضمون النص، فعلى سبيل المثال للكلمات غير الدالة في اللغة الإنجليزية هي ("a" و "the" و "is" و "are" وما إلى ذلك) تستخدم تقنية إزالة الكمات غير الدالة بعض الخوارزميات في النصوص لمعالجة اللغة الطبيعية؛ للقضاء على الكلمات التي المُستخدمة على نطاق واسعٍ، التي تحمل معلومات ذات فائدة قليلة (Ganesan، 2023).

على سبيل المثال، في سياق نظام البحث، إذا كنت تستعلم بحثيًا حول "ما هي تقنية معالجة اللغة الطبيعية؟"، فإنك تريد أن يركز نظام البحث على عرض المستندات، التي تتحدث عن "تقنية معالجة اللغة الطبيعية" بدلًا من المستندات التي تتحدث عن "ما هي" (Ganesan، 2023).

يمكن القيام بذلك من خلال الاحتفاظ بقائمة تشتمل على الكلمات المستبعدة (التي يمكن إعدادها يدويًا أو تلقائيًا)، ومنع تحليل جميع الكلمات الواردة داخل هذه القائمة، ففي المثال المذكور يمكن القضاء على الكلمات "ما هي"، مما يترك فقط الكلمات "معالجة اللغة الطبيعية"، وهذا يضمن تصنيف المستندات ذات الصلة بالموضوع بشكلٍ عالٍ في نتائج البحث الخاصة بك (Ganesan، 2023).

وتهدف هذه الخطوة إلى استبعاد الكلمات غير الدالة الشائعة الواردة داخل النص مثل (the، is، a، an) إلخ، مما يساعد عملية معالجة اللغط الطبيعية في التركيز على الأجزاء الأكثر جدوى من النص.

- **المرحلة الرابعة: تطبيع النص Normalization:**

تنطوي عملية تطبيع النص على تحويل النص إلى شكلٍ موحدٍ؛ لضمان الاتساق عبر مجموعة البيانات، ويتضمن ذلك مهامًا مثل تحويل جميع الأحرف إلى حالة صغيرة، وتوسيع

الاختصارات، فعلى سبيل المثال (تحويل"don't" إلى "do not")، كما يشمل التطبيع عمليتي التجذير والتصريف، التي سبق شرحها في المرحلة الأولى، فهذه الخطوة أساسية؛ لتقليل تعقيد النص والسماح بمعالجة أشكال مختلفة من نفس الكلمة، مما يسهل المقارنة والتحليل بشكل أفضل (Van Otten، 2023).

تعد عملية التطبيع بمثابة مهمة تنقيح بسيطة للنص ولكنها ذات تأثير كبير؛ لأنها تتضمن تحويل جميع النصوص إلى شكلٍ موحدٍ يسهل التعامل معه؛ للحفاظ على الاتساق، وضمان أن الخوارزميات لا تعامل الكلمات بشكلٍ مختلفٍ، عن طريق تحويل جميع النصوص إلى حروف صغيرة أو كبيرة لتوحيد النص، فهذه التقنية مفيدة عند التعامل مع بيانات نصية تحتوي على مزيج من الحروف الكبيرة والصغيرة، مثل التعرف على كلمتي "Apple" و "apple" ككلمة واحدة (Elsherif، 2024).

تجدر الإشارة هنا إلى طرق معالجة كتابة الأرقام والتواريخ، لا سيما أن الأرقام والتواريخ يمكن أن تضيف مستوى من التباين في النص، قد لا يكون مفيدًا لبعض مهام معالجة اللغة الطبيعية، ففي هذه الحالة يكون من المفيد إزالتها أو تحويلها إلى شكل موحد، كما أنه في بعض الحالات تحمل الأرقام والتواريخ معلومات حيوية، ويجب الاحتفاظ بها بطريقة تتماشى مع الأهداف التحليلية للنص، ولذلك فإن الأفضل تحويل الأرقام التواريخ إلى شكلٍ موحدٍ يمكن التعامل معه عند تنفيذ تقنية معالجة اللغة الطبيعية (Van Otten، 2023).

- **المرحلة الخامسة: تجزيء النص (Part of Speech (POS:**

يشكل تجزيء النص أساسَ فهم اللغة الطبيعية، حيث يتضمن تجزئ الكلام وضع تصنيفٍ مناسبٍ لكل كلمة في النص، مثل الاسم Noun، والفعل Verb، والصفة Adjective، وما إلى ذلك، فهذه العملية حاسمة للعديد من المهام الأخرى في مجال معالجة اللغة الطبيعية، حيث توفر معلومات أساسية عن البنية النحوية للجمل، مما يساعد في فهم معناها. (Elsherif، 2024).

ترجع أهمية تجزيء النص إلى تحديد أجزاء الكلام، وهو يعد أمرًا حيويًا لتحليل البنية النحوية للغة، مما يسهل فهم دلالة النص، كما يساعد تجزيء النص في توضيح المعاني وتفسير الكلمات التي يمكن أن تكون لها معان متعددة، ينتج عن هذا التجزيء مجموعة من

الكلمات التي تم تحليلها، وتحديد البنية النحوية للجمل والعبارات، مما يسهم في توافر معلومات أساسية لتطبيقات مختلفة في معالجة اللغة الطبيعية، بما في ذلك تحليل النصوص text analysis، والترجمة الآلية machine translation، واسترجاع المعلومات information retrieval (Mudadla، 2023).

يشتمل تجزيء النص على بعض المفاهيم الرئيسة، تلك المفاهيم عبارة عن **رموز قصيرة تمثل أجزاءً محددةً من النص، كما موضح بالجدول (1):**

جدول (1) رموز تجزيء النص في معالجة اللغة الطبيعية

نوع تجزيء النص	الاختصار
اسم Noun	(NN)
فعل Verb	(VB)
صفة Adjective	(JJ)
حال Adverb	(RB)
ضمير Pronoun	(PRP)
حرف جر Preposition	(IN)
أداة ربط Conjunction	(CC)
أداة تحديد Determiner	(DT)
صيغة تعجب Interjection	(UH)

- **المرحلة السادسة: تمثيل النص Text Representation:**

هذه المرحلة مسؤولة عن سد الفجوات بين النص المراد معالجته، وقدرة الآلة على تفسيره وتحليله واستخلاص المعاني المختلفة من ذلك النص، حيث تشمل هذه المرحلة مجموعة من التقنيات المسؤولة عن تحويل النص إلى أشكال عددية، تساعد الآلة من أداء مهام معقدة في مجال معالجة اللغة الطبيعية للنص المطلوب، وتلك التقنيات هي:

1- حقيبة الكلمات Bag of Words (BoW):

هو تمثيل مبسط يستخدم في معالجة اللغة الطبيعية واسترجاع المعلومات، حيث يتم تمثيل النص (مثل جملة أو مستند) على أنه حقيبة (مجموعة متعددة) من الكلمات، مع

الاحتفاظ بعدد تكرار كل كلمة داخل المستند، لذلك فهي تقنية قوية في معالجة اللغة الطبيعية؛ لقدرتها على استخراج المصطلحات الواردة داخل النص بسهولة ومرونة متجاهلة القواعد النحوية وترتيب الكلمات، كما يمكن تنفيذ تقنية حقيبة الكلمات بسهولة (Elsherif، 2024).

فعلى سبيل المثال: عند ورود الجملة التالية داخل النص:

"تعد تقنية حقيبة الكلمات من تقنيات تمثيل النص لمعالجة اللغة الطبيعية"

فإن تقنية حقيبة الكلمات تقوم بفرز الكلمات الواردة داخل المستندات كالآتي:

"تقنية"، "حقيبة"، "الكلمات"، " تقنيات"، " تمثيل"، " النص"، "معالجة"، "اللغة"، "الطبيعية".

فضلًا عن توضيح عدد تكرار كل مصطلح داخل الوثيقة، وقد أثبتت تلك التقنية كفاءتها في مختلف التطبيقات.

على الرغم من مزايا تلك التنقية وكفاءتها، لكن لديها بعض القيود التي تتضمن فقدان ترتيب الكلمات، وإغفال بعض المصطلحات، التي تتكون من أكثر من كلمة مثل مصطلح "معالجة اللغة الطبيعية" في المثال السابق، مما يؤدي إلى تحديات في التعرف على العلاقات الدلالية بشكلٍ كاملٍ، مما يترتب عنه بطبيعة الحال عدم الكفاءة الحسابية، لحل هذه المشكلة نعتمد على تقنية **(تردد الكلمة Term Frequency - تردد المستند العكسي Inverse Document Frequency).**

2- تردد الكلمة – تردد المستند العكسي (Term Frequency – Inverse Document Frequency):

تعالج تقنية (تردد الكلمة – تردد المستند العكسي) قيودَ تقنية حقيبة الكلمات، عن طريق تخصيص أوزان للكلمات الواردة داخل النص استنادًا إلى أهميتها في النص مقارنة بالمجموعة الكاملة من المستندات أو الوثائق التي تشمل نفس الكلمات، مما يساعد في تحديد الكلمات التي ليست فقط شائعة في مستند ما، ولكنها أيضًا مميزة لذلك المستند في سياق المجموعة الكاملة من المستندات (Eisenstein، 2019).

حيث يقصد بتردد الكلمة قياس عدد مرات تكرار المصطلح داخل المستند، ويحسب وزن المصطلح عن طريق قسمة عدد مرات ظهوره داخل المستند على إجمالي عدد المصطلحات داخل المستند، مما يعطي وزنًا كبيرًا للمصطلحات، التي تظهر بشكل متكرر في المستند، كما في المعادلة (1) التالية:

$$TF = \frac{\textit{no. of times term occurrences in a document}}{\textit{total number of terms in a document}}$$

أما تردد المستند العكسي، فهي تقيس أهمية المصطلح في مجموعة المستندات بأكملها، عن طريق قسمة العدد الإجمالي للمستندات على عدد المستندات، التي تحتوي على مصطلح ما، كما في المعادلة (2) التالية:

$$IDF = \frac{\textit{total number of documents}}{\textit{number of documents which are having term}}$$

ولحساب TF-IDF، يتم الحصول عليه عن طريق ضرب نسبة تكرار المصطلح الناتجة في المعادلة (1)، ونسبة تردد المستند العكسي الناتجة في المعادلة (2) كما في المعادلة التالية:

$$\textbf{TF-IDF} = TF * IDF$$

بناء على ما سبق، فقد تبين أن درجة أهمية المصطلح يتم تحديدها على حسب عدد مرات تكراره، فإذا كان ظهور مصطلح معين نادرًا داخل المستند، فذلك دلالة على عدم أهمية هذا المصطلح، وبناء على ذلك، يتم إعطاء المصطلح درجة منخفضة (Low Score)، أما إذا ورد المصطلح عدة مرات داخل الوثيقة، فهذا يدل على أن هذا المصطلح مهمٌ، وذو أثر على المستند، وعندها يعطى المصطلح درجة عالية (High Score).

3- **تضمين الكلمات Word Embedding:**

تُعرف بأنها عملية تحويل البيانات النصية إلى مقابلات عددية تستطيع الآلة فهمها من خلال استخدام خوارزميات التضمين، والتي تستخدم مخرجاتها كمدخل لخوارزمية التعلم الآلي Machine Learning المستخدمة في معالجة اللغة الطبيعية (Elsherif، 2024).

ترجع أهمية تقنية تضـــمين الكلمات إلى تطبيق التعلم الآلي على مجموعة بيانات كبيرة، حيث تجمع خوارزمية التضمين المدخلات، عن طريق وضع المدخلات المتشابهة لغويًا في مساحة التضـــمين الواحدة، وتجدر الإشـــارة إلى إمكانية تدريب النماذج على تضـــمين الكلمات وإعادة استخدامها لاحقًا.

وتعتمد تقنية تضمين الكلمات على خوارزمية تحويل الكلمات إلى متجهات Word2Vec، طور هذا الأســـلوب من قبل توماس ميكولوف Tomas Mikolov في عام 2013؛ لجعل تدريب البيانات في الشبكات العصبية مبني على التضمين، والذي يحقق أكثر كفاءة، ومنذ ذلك الحين أصبح الأسـلوب والمعيار الواقعي لتطوير تضـمين الكلمة، فهي عبارة عن مجموعة من النماذج التي تعمل على تمثيل الكلمة اعتمادًا على سياقها، .Rezaeinia et al، 2019)

وعند المقارنة بين التقنيات الثلاثة في مرحلة تمثيل النص، نجد أن لكل منهم خصـائص وســمات مختلفة عن الأخرى، يقدم الجدول (2) بعض الخصـــائص التي تميز كل تقنية عن الأخرى من حيث الكفاءة الحســـابية، والتمثيل الدلالي، وحجم البيانات التدريبية، وســـهولة التنفيذ، وسيناريوهات التطبيق (Durna، 2024).

جدول (2) التحليل المقارن لتقنيات تمثيل النص

عنصر المقارنة	BOW	TF-IDF	Word2Vec
الكفاءة الحسابية Computational Efficiency	معروفة ببساطتها وكفاءتها، وتعد مثالية لمعالجة سريعة في مجموعة من البيانات الصغيرة.	أكثر تطلبًا من BoW، بسبب حسابات الوزن، ولكنها متوازنة في القدرة على التعقيد.	عالية الكفاءة الحسابية، حيث تشمل تقنيات التعلم العميق، كما تقدم تحليلًا غنيًا لسياق النص.
التمثيل الدلالي Semantic Representation	يوفر تمثيلًا دلاليًا بسيطًا، حيث يفتقد السياق وترتيب الكلمات.	يقيس أهمية الكلمات عبر المستندات، ولكنه لا يلتقط السياق الكامل للنص.	يستطيع التعرف على المعاني الدلالية والعلاقات بشكل ممتاز، مما يقدم فهمًا دقيقًا.
حجم البيانات التدريبية Training Data	فعّالة مع البيانات المحدودة الصغيرة.	فعّالة مع البيانات المحدودة الصغيرة.	فعّالة لمجموعات البيانات الكبيرة.

عنصر المقارنة	BOW	TF-IDF	Word2Vec
سهولة التنفيذ Ease of Implementation	مباشرة، وتعد مثالية للمستخدمين المبتدئين في معالجة اللغة الطبيعية.	يتطلب فهم لغوي متوسط.	ينطوي على بعض الإعدادات المعقدة، والمعرفة العميقة بالشبكات العصبية Neural Networks.
سيناريوهات التطبيق Application Scenarios	مناسبة لتصنيف الوثائق والمستندات document classification، وتحليل المشاعر sentiment analysis. حيث لا تكون السياقات التفصيلية معقدة.	فعّالة في استرجاع المعلومات information retrieval، واستخراج الكلمات الرئيسية keyword extraction، وتجميع الوثائق document clustering، حيث تقدم توازنًا بين البساطة والارتباط في سياق النص.	مثالية للمهام التي تتطلب سياقات لغوية عميقة، مثل: الترجمة الآلية machine translation، والبحث السياقي contextual search، وتحليل المشاعر المتقدم advanced sentiment analysis.

تبين من الجدول السابق؛ أن لكل تقنية من التقنيات السابقة قوى مميزة في الكفاءة الحسابية، والتمثيل الدلالي للنص، وحجم البيانات التدريبية، ومتطلبات تنفيذ كل تقنية منهم، فضلًا عن السيناريوهات المثالية للتطبيق.

فقد تبين أنه يمكن تطبيق تقنية BoW وTF-IDF في السيناريوهات البسيطة التي تشمل مجموعات البيانات الصغيرة، نظرًا لبساطتهما وسهولة استخدامهم، على العكس، يُختار Word2Vecللمهام التي تتطلب فهمًا عميقًا وغنيًا لسياق النص المطلوب تحليله. وتجدر الإشارة إلى ضرورة النظر في المتطلبات الخاصة بالمشروع لاختيار النهج الأكثر فعالية.

7/2. خوارزميات معالجة اللغة الطبيعية.

لتنفيذ كل المراحل السابقة قد تحتاج إلى تطبيق بعض خوارزميات تقنية معالجة اللغة الطبيعية، وتحتاج خوارزميات التعلم الآلي Machine Learning لمعالجة اللغة الطبيعية إلى تدريب البيانات لمهام التعلم الخاضعة للإشراف Supervised Learning، مثل: التصنيف والتنبؤ، أو لمهام التعلم غير الخاضعة للإشراف Unsupervised Learning، مثل: التجميع.

فقد عرض كل من Aggarwal & Zhai (2012) مجموعة من الخوارزميات Algorithms التي يمكن تطبيقها عند إعداد معالجة للغة طبيعية في أي مجال، وهما:

1- **خوارزمية Naïve Bayes:** عبارة عن عائلة من الخوارزميات الاحتمالية، حيث تستخدم للتنبؤ بفئة النص.
2- **خوارزمية Linear Regression:** من الخوارزميات المعروفة المستخدمة في الاحصائيات، حيث تستخدم للتنبؤ ببعض القيم (Y) بناء على مجموعة من المزايا التي تم تدريبها (X).
3- **خوارزمية Support Vector Machine (SVM):** عبارة عن نموذج يستخدم لتمثيل أمثلة نصية كنقاط في مساحة متعددة الأبعاد، حيث يتم تعيين أمثلة للفئات المختلفة لمناطق مميزة داخل تلك المساحة، بعد ذلك؛ يتم تعيين فئة نصوص جديدة استنادًا إلى أوجه التشابه مع النصوص الموجودة والمناطق التي تم تعيينها لها.
4- **التعلم العميق Deep Learning:** عبارة عن مجموعة متنوعة من الخوارزميات التي تحاول مضاهاة العقل البشري، من خلال استخدام الشبكات العصبية الاصطناعية لمعالجة البيانات (Aggarwal & Zhai، 2012).

تجدر الإشارة إلى أن هناك بعض البرمجيات التي ظهرت مؤخرًا يمكنها تطبيق تلك الخوارزميات دون الاعتماد على الأكواد من خلال واجهة الاستخدام الرسومية Graphic User Interface (GUI)، مثل: (NLU "المستخدم في هذه الدراسة"، Orange).

8/2. تقنية التعرف على الكيانات المسماة Named Entity Recognition (NER).

تعد تقنية التعرف على الكيانات المسماة Named Entity Recognition (NER) عملية أساسية في استخراج وتحليل المعلومات من نصوص اللغة الطبيعية، حيث تساعد الآلة في التعرف على الكيانات المسماة من خلال تحديد المصطلحات وتصنيفها، كما أنها تفتح الأبواب أمام الآلات لفهم دلالات اللغة، مما يسهل التعامل مع العديد من النظم المتقدمة مثل: نظم استرجاع المعلومات information retrieval system، ونظم انشاء الرسوم البيانية المعرفية knowledge graph construction system، ونظم تحليل المحتوى content analysis system (Elsheirf، 2024).

ترتبط تقنية التعرف على الكيانات المسماة ارتباطًا وثيقًا مع تقنية معالجة اللغة الطبيعية، لذلك، مع تقدم تكنولوجيا معالجة اللغة الطبيعية، تستمر معها تقنية تعرف الكيانات المسماة في التطور، بالاعتماد على تقنية التعلم العميق Deep Learning، مما يقود الطريق نحو تحقيق دقة عالية ومرونة، فضلًا عن دور تقنية التعرف على الكيانات المسماة الحيوي في تعزيز الفجوة بين النصوص غير المهيكلة unstructured text، والفهم المنظم structured understanding (Lee، 2023).

تعتمد تقنية التعرف على الكيانات المسماة على العديد من التقنيات التي سبق ذكرها، مثل: التطبيع Normalization في عمليات توحيد النص standardizing the text، مما يجعل من السهل على النماذج تحديد وتصنيف الكيانات داخل النص بشكل صحيح، كما تعتمد على التأخيذ Tokenization في التعامل مع استخدام المسافات البيضاء داخل النص، والترقيم، والقواعد اللغوية داخل النص؛ لتحديد حدود الكلمات بدقة، مما يضمن عدم تقسيم الكيانات بشكل غير صحيح أو دمجها، وهو أمر حاسم للحفاظ على سلامة الكيان عبر النص Chowdhury، 2003).

بناء على ما سبق؛ يمكننا القول بأن تقنية التعرف على الكيانات المسماة هي عملية تحديد وتصنيف العناصر الرئيسية في النص إلى فئات محددة مسبقًا، مثل: (أسماء الأشخاص، والمؤسسات، والمواقع، والتواريخ، ...إلخ). لأن التعرف على الكيانات الواردة داخل النص أمر أساسي لاستخراج المعلومات المنظمة من البيانات النصية غير المنظمة.

1/8/2. منهجيات تقنية التعرف على الكيانات المسماة.

هناك مجموعة من المنهجيات المختلفة المستخدمة عند تطبيق تقنية التعرف على الكيانات المسماة، هي:

1. منهجيات قائمة على القواعد Rule-Based:

تعتمد تقنية التعرف على الكيانات المسماة في هذه الطريقة على مجموعة من القواعد اللغوية المصممة يدويًا والقواميس لتحديد الكيانات، قد تتضمن هذه القواعد بعض الأنماط للتعرف على الكيانات داخل النص، على سبيل المثال: كلمات السياق، مثل: ("رئيس" تأتي قبل

اســم الشــخص)، على الرغم من أن المنهجيات القائمة على القواعد يمكن أن تكون دقيقة للغاية في المجالات المحدودة والضــيقة، إلا أنها أقل مرونة وتتطلب جهدًا يدويًا كبيرًا لإنشــائها وصيانتها (Elsherif، 2024).

2. الأسلوب الاحصائي statistical Method:

تم الانتقال من القواعد اليدوية كما في الأسلوب السابق إلى الأسلوب الاحصائي المعتمد على مجموعة من الخوارزميات، مثل: : (Hidden Markov Models (HMMs، Conditional Random Fields (CRFs)، بغرض التعرف على المصطلحات داخل النص، والإشارة إلى الكيانات التي تنتمي إليها تلك المصــطلحات على أســاس الميزات المســتخرجة من النص، حيث يتم تدريب مثل هذه النماذج على مجموعات بيانات محددة مســبقًا، ثم يمكن تعميمها بشــكل جيد على النصــوص غير المرئية، على الرغم من أن أدائها الفعال، إلا أنها تعتمد بشــكل كبير على جودة البيانات التدريبية ومدى تمثيلها، فضــلًا عن أنها تتطلب تدريب حجم كبير من البيانات (Awan، 2023).

3. أسلوب تعلم الآلة Machine Learning Method:

تأخذ أساليب تعلم الآلة الأمور خطوة إلى الأمام عن طريق استخدام خوارزميات، مثل: Support Vector Machines (SVMs)، المســؤولة عن التنبؤ من البيانــات التي تم تــدريبها بالكيانات التي يمكن التعرف عليها داخل النص، يتميز هذا الأســلوب بالانتشــار الواســع في أنظمة التعرف على الكيانات المســماة الحديثة، بســبب قدرتها على التعامل مع مجموعات بيانات واســعة. ومع ذلك، فهي تحتاج إلى بيانات تدريبية كبيرة كما هو الحال في الأســلوب الاحصائي (Awan، 2023).

4. أسلوب التعلم العميق Deep Learning Methods:

أحدث ما توصــلت إليه تقنيات الذكاء الاصــطناعي بشــكل عام، وتقنية معالجة اللغة الطبيعية بشــكل خاص؛ هي أســاليب التعلم العميق، التي تســتغل قوة الشــبكات العصــبية Neural Networks، في اســتخراج الرؤى والمعاني من النصــوص المختلفة، ويمكن اســتخدام تلك التقنية في:

- تحليل محتوى المستندات، ورسائل البريد الإلكتروني.
- التلخيص التلقائي للمستندات أو المقالات الإخبارية.
- فهرسة العبارات الأساسية التي تدل على المشاعر، مثل التعليقات الإيجابية والسلبية على منشورات وسائل التواصل الاجتماعي (Awan، 2023).

9/2. تطبيقات تقنية معالجة اللغة الطبيعية في علم المكتبات والمعلومات.

يعد الهدف الأساسي من معالجة اللغة هو التواصل مع الحاسب بلغة طبيعية وصورة سهلة مما يسهل التعامل مع عددٍ كبيرٍ من المصطلحات، وقد ذكر كل من Taskin & Al. (2019) في الدراسة التي أعدوها بعض التطبيقات التي يمكن تنفيذها بالاعتماد على خوارزميات معالجة اللغة الطبيعية، ومنها:

1- **استرجاع النصوص Text Retrieval**: يقصد به الوصول إلى المعلومة المطلوبة داخل نص معين، حيث تسترجع النصوص من النص الأصلي المكتوب باللغة الطبيعية استجابة إلى سؤال أو استفسار يدخله المستخدم، ثم يبحث النظام عن نص له علاقة بالسؤال داخل مجموعة من النصوص، ولأن أغلب النصوص معقدة ومركبة فلم يحدث نجاح يذكر في هذا المجال إلا في مستويات بسيطة ومباشرة.

يمكن الاستفادة من تقنية معالجة اللغة الطبيعية في هذا الصدد داخل علم المكتبات والمعلومات، حيث يمكنها مساعدة الباحثين والمستفيدين من المكتبات ومراكز المعلومات في الوصول إلى أي جزء من فقرة داخل كتاب أو نص مقالي أو أي مصدر معلومات بدرجة كبيرة من الدقة، كما أنها تسمح باستخراج المعلومات في شكل كيانات مسماة، عن طريق تحديد أسماء الأشخاص، والمؤسسات، والتواريخ، والدول، وغيرها من الكيانات الأخرى داخل النصوص الكبيرة (Liddy، 2010).

وذلك يمثل موضوع تلك الدراسة، حيث تعتمد الدراسة على استرجاع بعض الكيانات، والكلمات الرئيسة، والفئات الموضوعية من النصوص، بالاعتماد على نموذج شُيد لتدريبه على مصطلحات مجال التربية وفهمها واسترجاعها عند ورودها داخل أي نص.

2- **الترجمة الآلية Machine Translation:** يعد هذا المجال من أولى المجالات، التي كان لها علاقة بمعالجة اللغة الطبيعية، حيث يعتمد على فهم اللغة، ونظرًا لأن الترجمة تتطلب مستوى أعلى من المعالجة والفهم، فلا تزال أغلب محاولات الترجمة في بداية الطريق، ولم تصل إلى مستوى الترجمة لمترجم محترف، يرجع ذلك إلى أن المشكلة الرئيسة في معالجة اللغة المراد ترجمتها تكمن في "كيف تتم عملية الفهم؟"، ومن المحاولات التي تمت لعلاج تلك المشكلة تتمثل في إيجاد لغة رياضية وسيطة تحتوي على كل المعاني والدلالات الواردة في اللغات الإنسانية؛ لأن المنطق والرياضيات كعلوم يمكن أن نعبر بها عن كل المعاني بصرف النظر عن اللغة نفسها (Taskin & Al، 2019).

3- **فهم وتوليد النصوص "فهم اللغة Language Understanding":** تتضح علاقة تطبيقات فهم اللغة بالتطبيق السابق (تطبيقات الترجمة الآلية) للغاية، ولكن الفرق أن الفهم يمكن أن ينقسم إلى عدة مستويات، تبدأ من لغة البرمجة العادية حتى فهم النصوص المعقدة، وعملية فهم النصوص لها عدة تطبيقات في مجال الذكاء الاصطناعي والنظم الخبيرة، حيث يعد أبسطها "تسهيل عملية الاستفسار والاستعلام" (Taskin & Al، 2019).

4- **التخليص Summarization:** تعتمد أنظمة التلخيص على استخدام بعض الأساليب اللغوية أو الإحصائية لاختيار الكلمات أو العبارات الأكثر أهمية من الجمل أو الفقرات في النصوص الكبيرة، ثم إنشاء ملخص لتمثيل النص (Blake، 2013).

10/2. مستقبل تقنية معالجة اللغة الطبيعية.

عند النظر إلى مستقبل تقنية معالجة اللغة الطبيعية نجد أنها تقدم إمكانيات لا حدود لها، بداية من تعزيز التعاون بين البشر والذكاء الاصطناعي، إلى كشف أسرار النصوص التاريخية، ودفع الاتصال العالمي، ولذلك يمكننا القول بإن تقنية معالجة اللغة الطبيعية تقف على طرف الموجة التالية من التطورات التكنولوجية. (Elsherif، 2024).

مع زيادة قدرات الحاسب الآلي وتطورها باستمرار، يمكننا التوقع بتقنيات معالجة اللغة الطبيعية، التي تكون أكثر تطورًا وشمولًا وقدرة على التعامل مع التفاصيل المعقدة، فضلًا عن

التوقع بظهور الكثير من البرمجيات ذات الواجهات الرسومية Graphic User Interface، التي ستسمح باستخدام تقنيات معالجة اللغة الطبيعية دون الحاجة استخدام أكواد Encoding، مما يساعد المستخدمين المبتدئين في تطبيق تقنيات معالجة اللغة الطبيعية بسهولة (Lee، 2023).

عند النظر إلى تأثيرات مجال معالجة اللغة الطبيعية العميقة على المستقبل، نكتشف أن هناك بعض الإمكانيات الهائلة لمعالجة اللغة الطبيعية القادرة على تحويل عالمنا، مما يجعله أكثر ارتباطًا وتمكنًا وفهمًا، حيث يعد مجال معالجة اللغة الطبيعية ميدانًا في قلب التجربة البشرية والابتكار التكنولوجي؛ لأن هذا المجال يحتل مكانة فريدة في مفترق طرق الذكاء الاصطناعي (Artificial Intelligence AI)، وعلم اللغة linguistics، وعلم البيانات data science، مما جعل مجال معالجة اللغة الطبيعية قادرًا على تقديم رؤى وقدرات محورية تعيد تشكيل هذه التخصصات (Elsherif، 2024).

11/2. تحديات تقنية معالجة اللغة الطبيعية.

عند الحديث عن التحديات والصعوبات التي تواجه مجال معالجة اللغة الطبيعية في هذا الصدد، نجد أن هناك بعض الصعوبات في معالجة وتحليل حجم كبير من البيانات النصية، لا سيما بالنسبة للغات ذات الهياكل النحوية المعقدة، كما تعد التوافقية Interoperability تحديًا آخر، حيث تحتاج نظم معالجة اللغة الطبيعية إلى التكامل والتوافق مع البيئات البرمجية الحالية والمعايير المتاحة، لذلك تحتاج تقنيات معالجة اللغة الطبيعية من النشر والاستخدام بشكلٍ فعالٍ عبر منصات وتطبيقات مختلفة (Khurana et al.، 2022).

هذه التحديات تؤكد الدقة المطلوبة في نماذج وخوارزميات معالجة اللغة الطبيعية؛ لمعالجة وتوليد اللغة الطبيعية بدقة، فعلى سبيل المثال عند تطبيق تقنية معالجة اللغة الطبيعية لتحليل محتوى مستند ما دون تنسيق مستند، قد يكون صعبًا تحقيق نتائج دقيقة باستمرار عند استخدام معالجة نصية حرة لاستخراج حقائق محددة من هذا المستند (Elsherif، 2024).

كما أن هناك بعض التحديات في معالجة اللغات الطبيعية في العديد من اللغات، لا سيما التحديات المرتبطة باللغة العربية نفسها ولهجاتها عند معالجة اللغة الطبيعية، حيث توجد بعض المشكلات المرتبطة بغموض المصطلحات Ambiguity المتمثلة في أن يكون لمصطلحٍ واحدٍ أكثر من معنى، ومشكلات فقدان الدقة Imprecision، حيث يرجع ذلك إلى أن معظم الأشخاص غالبًا ما يعبرون عن أفكارهم باستخدام مفردات أو كلمات غير دقيقة، فضلًا عن مشكلات فقدان الاكتمال Incompleteness، غالبًا ما يغفل بعض الأشخاص في حديثهم عن الكثير من التفاصيل دون فقدان المعنى المراد توصيله.

ويرجع ذلك إلى أن الأشخاص يمكنهم عن طريق خبرتهم السابقة فهم ما يعنيه المتحدث دون الإسهاب في الشرح، مثل هذه المشكلات التي تواجه الباحثين في مجال الذكاء الاصطناعي، وتحديدًا معالجة اللغة الطبيعية، لا يوجد لها حلول كاملة حتى الآن، ولكن الجهود التي تبذل في هذا المجال تعمل على حل بعضها (shelf، 2024).

12/2. الخلاصة.

استعرضت الدراسة تقنية معالجة اللغة الطبيعية من حيث التعريف، والتاريخ، فقد قدمت الدراسة عرضًا تاريخياً لتقنية معالجة اللغة الطبيعية من خلال (5) مراحل، تبدأ من قبل ستينيات القرن الماضي حتى الآن، موضحًا المصطلحات الخمسة الأساسية المستخدمة في مجال معالجة اللغة الطبيعية (الجملة Sentence، الكلمة Utterance، الترميز Tokenization، التجذير Stemming، التصريف Lemmatization)، فضلًا عن العناصر الثلاثة التكوينية لتقنية معالجة اللغة الطبيعية (فهم اللغة الطبيعية NLU، توليد اللغة الطبيعية NLG، اكتساب المعرفة والاستدلال (KAI).

كما تناولت الدراسة المستويات اللغوية الستة في مجال معالجة اللغة الطبيعية (الصوتيات Phonetics، علم الصوتيات Phonology، الصرف Morphology، النحو Syntax، المعاني Semantics، البرجماتية pragmatics)، مع توضيح المراحل التي تمر بها تقنية معالجة اللغة الطبيعية (المرحلة الأولى: المعالجة المسبقة للنص، المرحلة الثانية: التأخيذ، المرحلة الثالثة: إزالة الكلمات غير الدالة، المرحلة الرابعة: تطبيع النص، المرحلة الخامسة: تجزيء النص، المرحلة السادسة: تمثيل النص).

كما أشـارت الدراسـة إلى الخوارزميات المسـتخدمة عند تطبيق تقنية معالجة اللغة الطبيعية (Naïve Bayes، Linear Regression، Support Vector Machine (SVM))، وقـد أشـارت الدراسـة إلى تقنية التعرف على الكيانات المسـماة باعتبارها العملية الأسـاسـية في اسـتخراج وتحليل المعلومات من نصوص اللغة الطبيعية، وتصنيف المصطلحات إلى كيانات.

وأخيرًا، تناولت الدراسـة بعض تطبيقات معالجة اللغة الطبيعية التي يمكن تطبقها في مجال علم المكتبات والمعلومات، مثل: (اسـترجاع النصـوص، والترجمة الآلية، وفهم وتوليد النصـوص "فهم اللغة"، والتلخيص)، فضـلًا عن تقديم الدراسـة لمسـتقبل تقنية معالجة اللغة الطبيعية، والتحديات والصعوبات التي تواجه مجال معالجة اللغة الطبيعية في هذا الصدد.

المراجع

أولًا: المراجع العربية:

السلمي، عفاف سفر. (يوليو 2017). تطبيقات الذكاء الاصطناعي لاسترجاع المعلومات في جوجل. *مجلة دراسات المعلومات – جمعية المكتبات والمعلومات السعودية. ع19. ص: 103-124.*

ثانيًا: المراجع الأجنبية:

Awan, A. A. (2023, September 13). What is Named Entity Recognition (NER)? Methods, Use Cases, and Challenges. https://www.datacamp.com/blog/what-is-named-entity-recognition-ner

Bender, E. M. (2013) Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax (Synthesis Lectures on Human Language Technologies). Morgan & Claypool Publishers.

Blake, C. (2013). Text mining. Annual Review of Information Science and Technology, 45(1), 121-125.

Brank, J., Mladenič, D., & Grobelnik, M. (2016). Feature construction in text mining. In Springer eBooks (pp. 1–6). https://doi.org/10.1007/978-1-4899-7502-7_100-1

Chowdhury, G.G. (2003). Natural language processing. Annual Review of Information Science and Technology, 37(1), 51-89.

Daniel. (2023, October 30). Natural Language Processing (NLP): definition and principles. Data Science Courses | DataScientest. https://datascientest.com/en/natural-language-processing-definition-and-principles

Durna, M. B. (2024, January 11). Text Representation Techniques - Merve Bayram Durna - Medium. Medium. https://medium.com/@mervebdurna/text-representation-techniques-d40741eb0916

Eisenstein, J. (2019) Introduction to Natural Language Processing (Adaptive Computation and Machine Learning series). The MIT Press.

Elsherif, H. M. (2024). Natural Language Processing (NLP): The Complete Guide.

Fingent. (2023, July 18). Knowledge representation models in artificial intelligence. Medium. https://fingent.medium.com/knowledge-representation-models-in-artificial-intelligence-e33d180ef7be#:~:text=Knowledge%20Representation%20is%20a%20field,handle'%20real%2Dlife%20tasks

Ganesan, K. (2023, March 12). What are Stop Words? Opinosis Analytics. https://www.opinosis-analytics.com/knowledge-base/stop-words-explained/#:~:text=Stop%20words%20are%20a%20set,carry%20very%20little%20useful%20information.

Green, B., Wolf, A., Chomsky, C. and Laughery, K. (1961). BASEBALL: an automatic questionanswerer. In Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference (IRE-AIEE-ACM '61 (Western)). Association for Computing Machinery, New York, NY, USA, 219–224.

Hausser, R. (2014) Foundations of Computational Linguistics: Human-Computer Communication in Natural Language (3rd edition). Springer.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Lee, R. S. T. (2023). Natural language processing: A Textbook with Python Implementations. Springer.

Lee, R. (2024). Natural language processing. https://doi.org/10.1007/978-981-99-1999-4.

Levine-Clark, M., & Carter, T. (2021). ALA glossary of library and information science (4th ed., p. 154). American Library Association.

Liddy, E.D. (2010). Natural language processing. In Encyclopedia of Library and Information Sciences, Third Edition (pp. 3864-3873). New York: Taylor and Francis.

Mudadla, S. (2023, November 10). What is Parts of Speech (POS) Tagging Natural Language Processing?In what kind of applications we can use Parts of Speech (POS) Tagging in Natural Language Processing. Medium. https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186#:~:text=Part%2Dof%2DSpeech%20(POS,grammatical%20roles%20of%20individual%20words.

Murel, J., & Kavlakoglu, E. (2023, December 10). What are stemming and lemmatization? | IBM. Retrieved April 2, 2024, from https://www.ibm.com/topics/stemming-lemmatization

Natural Language Processing. (2017). Oxford Living Dictionaries. Erişim adresi: https://en.oxforddictionaries.com/definition/natural_language_processing.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. Expert Systems With Applications, 117, 139–147. https://doi.org/10.1016/j.eswa.2018.08.044

Santilal, U. (2020) Natural Language Processing: NLP & its History (Kindle edition). Amazon.com.

Shelf. (2024, March 21). Challenges and considerations in natural language Processing. Shelf. https://shelf.io/blog/challenges-and-considerations-in-nlp/

Taskin, Z., & Al, U. (2019). Natural language processing applications in library and information science. Online Information Review, 43(4), 676–690. https://doi.org/10.1108/oir-07-2018-0217

Turing, A. (1950) Computing Machinery and Intelligence. Mind, LIX (236): 433–460.

Van Otten, N. (2023, November 1). How to use text normalization techniques in NLP with Python [9 ways]. Spot Intelligence. https://spotintelligence.com/2023/01/25/text-normalization-techniques-nlp/#:~:text=Text%20normalization%20is%20a%20key,removal%2C%20stemming%2C%20and%20lemmatization.

Arab International Journal of Information Technology & Data
Vol. 4, No. 4 October - December 2024

Natural Language Processing Techniques for Research and Retrieval Purposes in Library and Information Science

Dr. Mostafa M. I. Y. El-Helaly
Cairo University, Faculty of Arts
Department of Library, Archives and Information Technology
mostafaelhelalyy@cu.edu.eg

Supervised by
Prof. Osama A. G. Alqlsh
Cairo University, Faculty of Arts
Department of Library, Archives and Information Technology
alqlsh@yahoo.com

Abstract:

Natural Language Processing (NLP) is a branch of artificial intelligence technologies that has made interacting with computers more akin to natural language. This study aimed to define NLP, present its history from the sixties to the present, clarify the fundamental terms used in the field of NLP, as well as identify the constituent elements of NLP technology, the linguistic levels in the field of NLP, the stages involved in NLP technology, and the applications of NLP in library and information science. The study relied on a descriptive-analytical approach in reviewing the intellectual production related to the field of NLP, based on available foreign databases on the Egyptian Knowledge Bank (EKB). The study reached several conclusions, notably anticipating the emergence of numerous graphical user interface software that will allow the use of NLP techniques without the need for encoding, thereby facilitating beginners in applying NLP techniques easily without relying on algorithms.

Keywords: Natural Language Processing; Natural Language Understanding; Artificial Intelligence; Information Retrieval; Text Retrieval.