Classification Software Engineering Documents Based on Hybrid Model

Nada N. Saleem **Rasha Gh. Saeed**
Nada_N_S@uomosul.edu.iq
College of Computer Sciences and Mathematics
University of Mosul, Mosul, Iraq

Received on: 16/07/2018 **Accepted on: 27/08/2018**

ABSTRACT

Care about automated documents classification has increased since the appearance of the digital documents and the wide diffusion of Internet. In the 1990's, the computer performance has greatly improved and has led to the methods of machine learning to establish automated classifiers. These methods have achieved good speed and classification's accuracy and researchers still investigate in this field to accomplish more accuracy and less time. Artificial immunologic systems have shown high performance in such as data clustering and anomaly detection which can be ascribed to the nature of the immunologic system in protecting the body.

Some of the present methods and ways used in the training process of the document classification are time consuming and others have less accuracy rate concerned with the classification of the related document as software engineering document classes. For these reasons, this research deals with the study of Natural Immune System and using the dynamic process of the Adaptive Immune System work by hybridization Negative Selection (NS) and Positive Selection (PS) techniques and to propose a hybrid model called the Hybrid Positive Negative Selection Model (HPNS) to classify Software Engineering documents as they comprise information related to developing the software systems, that makes it easy for the software engineer who works in maintenance.

HPNS has high classification's speed and accuracy besides easy and flexible use by designing interfaces that make it easy for the user to deal with the system. In order to improve the quality and the efficiency of HPNS method, it was compared to one of the best and well-known methods of classification referred to as, Naive Bayes(NB). After conducting several experiments on a various group of software engineering documents, evaluations results have shown that the accuracy of the Adaptive immunologic method (HPNS) has reached (HPNS) (95%), whereas Naïve classification method has reached (90 %) with training and classification speed that doesn't exceed one minute. This shows the feasibility of using the algorithms of AIS systems in the field of information recovery and documents classification. This system was built and programmed in Java language and was implemented under an operating system environment Microsoft Windows7.

Keyword: Artificial immunologic systems, Negative Selection, Positive Selection, Software Engineering Documents, Documents Classification.

تصنيف وثائق هندسة البرمجيات بالاعتماد على أنموذج هجين

ندى نعمت سليم **رشا غانم سعيد**
كلية علوم الحاسوب والرياضيات
جامعة الموصل، الموصل، العراق

تاريخ استلام البحث: 2018\07\16 **تاريخ قبول البحث: 2018\08\27**

الملخص

تزايد الاهتمام بالتصنيف الآلي للوثائق منذ ظهور الوثائق الرقمية والانتشار الواسع للانترنيت، وفي التسعينات تحسن أداء الحاسوب على نحو كبير والذي قاد إلى استخدام طرائق تعليم الآلة لتكوين مصنفات آلية حققت دقة تصنيف وبسرعة جيدة، وما زال الباحثون يبحثون في هذا المجال لتحقيق دقة اكبر ووقت اقل. وقد أظهرت الأنظمة المناعية الاصطناعية أداءً عالياً في مهمة التصنيف مثل عنقدة البيانات وكشف الشذوذ وذلك يعود لطبيعة النظام المناعي في حماية الجسم.

إن بعض الطرائق والأساليب الحالية تحتاج إلى وقت في عملية التدريب لتصنيف الوثائق والبعض الآخر من الأساليب تكون دقة تصنيفها للوثائق قليلة فيما يتعلق بالوثائق المتقاربة الأصناف مثل أصناف وثائق هندسة البرمجيات, لذلك تناول هذا البحث دراسة النظام المناعي الطبيعي واستخدام ديناميكية عمل النظام المناعي التكيفي وذلك بتهجين تقانة الانتقاء السلبي (Negative Selection (NS وتقانة الانتقاء الايجابي (Positive Selection (PS، واقتراح أنموذج مهجن المسمى أنموذج الانتقاء الإيجابي والسلبي Hybrid Positive Negative Selection Model(HPNS) لتصنيف وثائق هندسة البرمجيات لما تحويه من معلومات متعلقة بتطوير الأنظمة البرمجية التي تسهل العمل على نحو كبير على مهندس البرمجيات الذي يقوم بالتطوير والصيانة.

تمتلك طريقة التصنيف (HPNS)دقة وسرعة عالية في التصنيف حتى فيما يتعلق بتصنيف الوثائق المتقاربة الأصناف ولبرهنة فاعلية وجودة طريقة(HPNS) قورنت بأحد أفضل وأشهر طرائق التصنيف المستخدمة وهي طريقة Naïve Bayes (NB)، وبعد إجراء التجارب على مجموعة متنوعة من وثائق هندسة البرمجيات ، أظهرت نتائج التقييم أن دقة طريقة تصنيف المناعة التكيفية(95%) (HPNS)، في حين أن دقة تصنيف 90%) ((NB), وبمعدل سرعة تدريب وتصنيف لا يتجاوز الدقيقة الواحدة, مما يظهر جدوى استخدام خوارزميات أنظمة ((AIS Artificial Immune System في حقل استرجاع المعلومات وتصنيف الوثائق. بني هذا الأنموذج وبرمجته بلغة جافا وتنفيذه تحت بيئة نظام تشغيل مايكروسوفت ويندوز Microsoft windows7.

الكلمات المفتاحية: الأنظمة المناعية الاصطناعية، الانتقاء السلبي, الانتقاء الإيجابي، تصنيف الوثائق، وثائق هندسة البرمجيات.

1. مقدمة

أصبح تصنيف الوثائق الآلي مجال بحث وتطبيق مهم منذ ظهور الوثائق الرقمية؛ إذ يعد تصنيف النص ضرورياً جداً نتيجة الحجم الهائل للوثائق النصية التي نتعامل معها يومياً, ومع تزايد النمو المستمر للشبكة العنكبوتية (www)، فإن تحديد المعلومات ذات العلاقة بالبحث أصبح عملية صعبة جداً. ولاسترجاع المعلومات بأقل وقت ممكن وبأعلى درجة من العلاقة جاء دور التصنيف الآلي؛ إذ إن تصنيف الشبكة الآلي يساعد في استرجاع المعلومات بشكل أفضل [10], استخدمت الوثائق لتوثيق أنواع متعددة من المعلومات ومنها المعلومات الخاصة بكل مرحلة من مراحل هندسة البرمجيات؛ إذ تلعب وثائق مراحل التطوير دوراً مهماً في بناء الأنظمة البرمجية، ويشكل التوثيق وسيلة لتقديم الرؤية ضمن عملية البرمجيات، ويعد دور الاتصال أساسيا لتطوير البرمجيات[18]. يعد امتلاك مخزن للمعرفة من العوامل المهمة في نجاح أي نوع من الأعمال التقانية الكبيرة ومن الضروري تسجيل تلك المعرفة في الوثائق؛ إذ تكون الوثائق بصيغة الكترونية ممكن البحث عنها ومشاركتها واسترجاع الناس بفاعلية اكبر للمعلومات الذين بحاجة لها داخلياً وخارجياً[8]. واتجه الباحثون إلى استخدام المفاهيم الذكائية لحل مشكلات استرجاع المعلومات والتصنيف وقد حققوا انجازات في ذلك، ومن المفاهيم الحديثة في حقل الذكاء الاصطناعي الأنظمة المناعية الاصطناعية التي استخدمت في البحث وبناء مصنف لتصنيف الوثائق لمراحل هندسة البرمجيات.

2. الدراسات السابقة

حاول الباحثون في السنوات الماضية إيجاد حلول واقتراحات كثيرة لتصنيف الوثائق بطرائق سهلة، سريعة، ذات كفاءة عالية وكلفة قليلة، وذلك باستخدام تقانات تعليم الآلة التي كان لها إمكانية في تحسين أداء المصنفات، ومن هذه البحوث:

1. في عام 1997 قدم الباحثان Pazzani و Billsus بحثاً عن عدد من خوارزميات التعلم المختلفة, واختبار أداء ودقة مصنف (Naïve Bayes (NB ، والمجاور الأقرب، واشجار القرار والشبكات العصبية ووجد أن أداء مصنف NB الأفضل بصورة عامة. وأيضاً بين دور اختيار الخواص في زيادة دقة المصنف وتقليل خطأ التصنيف [31].

2. في عام 1998 قدم الباحث Joachims بحثاً عن الة دعم المتجه (Support Vector Machine (SVM بوصفها طريقة لتصنيف الوثائق؛ إذ أثبتت هذه الطريقة كفاءتها بالتصنيف إلا أنها تعاني من التعقيد، ويقع التعقيد في تضبيط المعاملات واختيار Kernal [17] .

3. في عام 1999 قدم الباحثان Cohen وSinger بحثاً لطرائق التعليم لتصنيف النص واستخدما أشجار القرار وبينَا قابليتها في التصنيف؛ إذ تمتاز بسهولة الفهم وتقلل تعقيد المشكلة إلا أن أشجار القرار تتطلب وقت تدريب أطول، كما أن الوثيقة تُمثل في هيكلية شجرة، وفي هذه الحالة عند حصول خطأ في مستوى عالٍ فإن أي شجرة فرعية ستكون خاطئة[4].

4. في عام 2002 قدم الباحث Sebastiani بحثاً عن طرائق التصنيف واستخدام الشبكة العصبية وامكانيتها في تصنيف الوثائق، وبين ان الطريقة المثالية لتدريب الشبكة العصبية هي الانتشار الخلفي back propagation؛ إذ عند حصول خطأ بالتصنيف يتم ارجاع الخطأ ونشره ثم تغير المعاملات للشبكة ثم يقل الخطأ, قدمت الشبكة العصبية نتائج جيدة إلا أن تدريب الشبكة يكون بطيئاً [35].

5. في عام 2002 قدم الباحثان Zuben و Decastro بحثاً يتناول التصنيف باستخدام خوارزمية الانتقاء النسيلي، وذلك بتوليد كرات التمييز الاصطناعية(Artificial Recognition Balls(ARBs التي تستجيب للمستضدات (مجموعة متجهات التدريب)، الكرات الأكثر استجابة تُخزن في مخزن خلايا الذاكرة وتُكون أداة التصنيف، استخدمت الطفرة في الخوارزمية مما جعل ARBs غير مماثلة لمتجهات التدريب من جهة، ومن جهة أخرى تكون متشابهة كفاية لتدريب متجهات أخرى, بيّن البحث قدرة نظام المناعة على التعلم وكفاءته في التصنيف، وأن عدد خلايا الذاكرة المتكونة عادة يكون نصف عدد خلايا التدريب[5].

6. في عام 2005 قدم الباحثون Paab و Nurnberger و Hotho بحثاً عن استخدام مصنف المجاور الأقرب في تصنيف الوثائق الذي يكون فعالاً إلا أن وقت التصنيف طويل جداً واختيار القيمة المثالية لـ K(عدد المجاورات) تعد قضية صعبة؛ إذ إن اختيار قيمة واحدة لـ K تلائم كل التطبيقات أمراً مستحيلاً [14].

7. في عام 2007 قدم الباحثان Romero وNino طريقة لاستخلاص الكلمات ذات الدلالة من الوثائق النصية بالاعتماد على نظام المناعة؛ إذ تعتمد على خلفية رياضية التي تُعرف بطريقة قياسية: التفاعل بين الأجسام المضادة في الشبكة المناعية الاصطناعية، مثل هذا التفاعل يؤدي إلى استخلاص الكلمات ذات الدلالة(التي تنفع في التصنيف) ومن خلال الكلمات المستخلصة من الوثائق النصية من الممكن تمثيل المعرفة لكل صنف، فضلاً عن إمكانية تكوين قائمة من stopword من خلال الاجسام المضادة الأقل تحفيزاً وحذفها من الشبكة(حذف الكلمات ذات التحفيز القليل من الشبكة)[34].

8. في عام 2008 قدم الباحث Tan بحثاً يتناول استخدام طريقة Centroid لتصنيف الوثائق الذي يعد من الطرائق الإرشادية الشائعة في التصنيف والمعروف ببساطته, قدم هذا الباحث تحسيناً على المصنف بإضافة معامل ثابت يدعى نسبة التعلم والذي حسن من كفاءة المصنف، إلا أنه يعاني من عدم الملاءمة الناتجة من فرضيته؛ وهي أن الوثيقة المعطاة يجب تعيينها لصنف معين إذا كان التشابه لمركز صنف معين كبير، وعادة يكون هذا الافتراض غير صحيح في الحالات العملية[37].

9. في عام 2012 قدم الباحثان . Pawar وGawande بحثا يقدم مقارنة بين انواع مختلفة لطرق التصنيف: Naïve Bayes ، المجاور الاقرب، اشجار القرار ,الشبكات العصبية ,خوارزمية Rocchio وآلة دعم المتجه Support Vector Machine (SVM) ودمج هذه الطرق كطرق هجينة مثل تهجين الشبكة العصبية مع اشجار القرار, فضلا عن تقنيات اختيار الخواص ودورها في زيادة دقة المصنف.

10. في عام 2015 قدم الباحث Onan بحثا يتناول النظام المناعي الاصطناعي وفاعليته في تصنيف صفحات الويب واستخدام خوارزمية immunes_1 و immunes99, ثم مقارنتهما بتقنيتي التعلم C4.5 مصنف شجرة القرار و مصنف Naïve Bayes

ومن خلال التجارب اثبت ان الانظمة المناعية الاصطناعية تحقق اداءاً افضل لتصنيف صفحات الويب.

11. في عام 2015 قدم الباحثان Saranya و Thenmozhi بحثا يتناول تمثيل معنى النص لتقليل الخواص باستخدام word Net ثم تطبيق الة دعم المتجه Support Vector Machine (SVM) لتصنيف الوثائق, ومن ثم مقارنة اداء الخواص الاصلية بالخواص التي تم تقليلها, وقد اظهرت النتائج ان الاداء الذي تم الحصول عليه بواسطة البيانات المقللة افضل من البيانات الاصلية.

12. في عام 2017 قدم الباحثان Semberecki و Maciejewski بحثا عن استخدام الشبكة العصبية مع وحدات (LSTM)(Long Short Term Memory), واختبار عدد من طرق تمثيل متجه الخواص ومن خلال التجارب اثبت ان النهج القائم على شبكة (LSTM) مع الوثائق الممثلة بمتجه متسلسلات من الكلمات المشفرة يفوق الوثائق الممثلة بمتجه خواص تكرارات الكلمة.

3. جهاز المناعة الطبيعي

لقد تمت دراسة جهاز المناعة عند الإنسان بصورة جيدة لما يزيد عن المئة عام, ولكن لا يزال هذا الجهاز لم يفهم بشكل كامل. فجهاز المناعة هو نظام دفاعي تطور ليحمي المضيف من الكائنات الممرضة Pathogens (الكائنات الحية المجهرية المؤذية مثل البكتريا والفايروسات والطفيليات) [11]؛ إذ يتكون الجهاز من الخلايا المتخصصة المتنوعة التي تنتشر وتراقب الجسم وكذلك جزيئات خلوية إضافية متنوعة وتنظيمات مناعية تعمل على توفير بيئة للخلايا المناعية لتتفاعل وتنمو وتستجيب[15]. تنشأ الخلايا المناعية في نخاع العظم والتوتة وعند نضوجها تهاجر إلى الأنسجة متنقلة عبر الأوعية الدموية واللمفاوية.

إن الوظيفة الرئيسة لتلك الخلايا هي تمييز وجود العناصر الغريبة في الجسم وتعمل على الاستجابة لتتخلص منها أو لتقضي على تأثير الداخلين الأجانب[20]. تدعى المواد التي يمكن أن تحفز استجابات معينة في الجهاز المناعي عادة بالمُستضدات Antigen(Ag) (الكائنات الممرضة تعمل عادةً مستضدات) [27].

إن نظام المناعة الطبيعي يتكون من خط دفاعي ذي طبقتين تعرفان بأنهما جهاز المناعة الفطري وجهاز المناعة التكيفي. يعتمد كلا النظامين على فعالية خلايا الدم البيضاء(الكريضات)[6].

1.3 جهاز المناعة الفطري

جهاز المناعة الفطري (غير المكتسب) هو الخط الأول للدفاع الذي يعمل على تقديم استجابة فورية ولكنها تكون غير محددة ضد العامل المسبب للمرض المهاجم مثل البكتريا والفايروسات[39].

2.3 جهاز المناعة التكيفي

يشار إلى جهاز المناعة التكيفي أيضاً بأنه جهاز المناعة المكتسبة؛ إذ تطلق عليه هذه التسمية؛ لأنه مسؤول عن تخصيص دفاع للكائن الحي المضيف الذي سيواجه الخطر, وذلك بالاعتماد على العامل الممرض المحدد،على عكس جهاز المناعة الفطري، يوجد جهاز المناعة المكتسبة بالفقريّات فقط (الحيوانات ذات العمود الفقري)؛ إذ يحتفظ الجهاز بذاكرة للهجمات التي واجهها[2]. إن جهاز المناعة التكيفي هو خط الدفاع الثاني الذي يتوسط استجابة محددة ومتأخرة [39].

ينظم جهاز المناعة المكتسبة حول صنفين من الخلايا: الخلاياالتائية T-Cell والخلايالبائية B-Cell، يعد صنف خلايا جهاز المناعة الفطري أكثر تعدداً، وبضمنها الخلايا القاتلة (NK) Natural killer الخلايا التغصنية (DCs) والخلايا البَلْعَميةِ الكبيرة. تتفرع الخلايا ضمن هذه الأصناف إلى أنواع مختلفة، مثل خلية Tالبسيطة أو المساعدة، وخلية DC غير الناضجة أو شبه الناضجة أو الناضجة[38].

4. نظريات NIS

توجد نظريات مختلفة عن دراسة علم المناعة فيما يتعلق بالسلوك الوظيفي والتنظيمي مابين الخلايا اللمفاوية عند الاستجابة إلى المُستضد المهاجم الذي تواجهه، تضم هذه النظريات[7]:

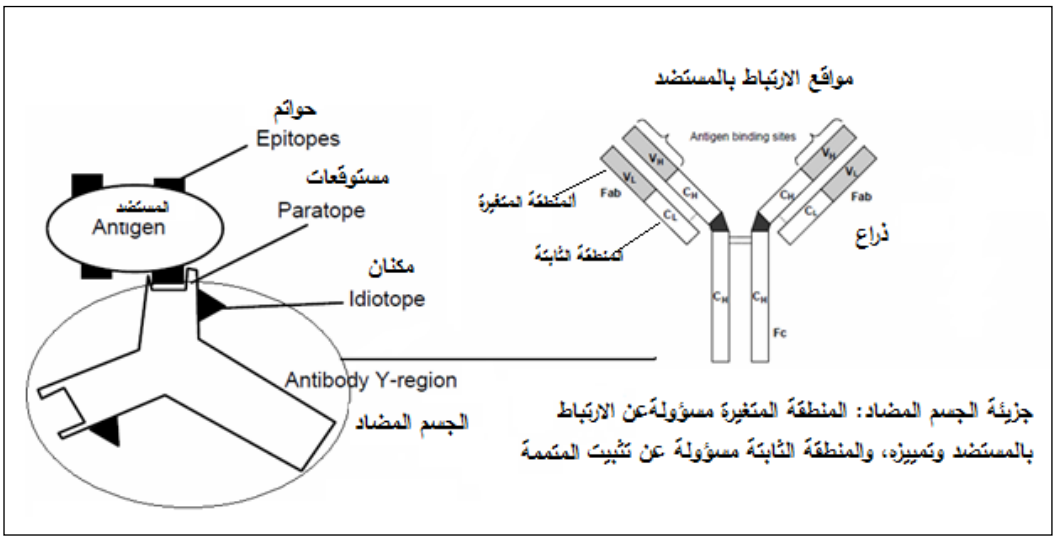
1- النظرية التقليدية.
2- نظرية الانتقاء النَسيليّ.
3- نظرية شبكة المناعة.
4- نظرية الخطر.

1.4 النظرية التقليدية

إن النظرية التقليدية لجهاز المناعة هي إن جهاز المناعة يميز مابين ما هو طبيعي (ذاتي) وغريب (غير ذاتي أو مستضد) في الجسم [15]؛ إذ يؤدي تمييز المُستضدات إلى خلق الخلايا المُفْعّلة المتخصصة التي تعمل على خمول وتدمير هذه المُستضدات.

يتكون جهاز المناعة الطبيعي من الخلايا اللمفاوية وأعضاء لمفاوية، هذه الأعضاء هي اللوزتان، الغدد، التوتة (الغدة الصعترية)، العقد اللمفاوية، الطحال، بقع باير patche's peyer، الزائدة الدودية، الأوعية اللمفاوية ونخاع العظم. إن الأعضاء اللمفاوية مسؤولة عن عمل الخلايا اللمفاوية ونموها وتطورها في جهاز المناعة، تستخدم الخلايا اللمفاوية للكشف عن أي مستضدات موجودة في الجسم، ويعمل جهاز المناعة على مبدأ نظام تمييز الأنماط، أي تمييز الأنماط غير الذاتية عن الأنماط الذاتية[7]. لقد عرف Burnet [3] النظرية التقليدية

الأولى بأن الجهاز المناعي يتكون من الخلايا البائية والخلايا التائية القاتلة مع مستقبلات مستضد معين. إذ تطلق المُستضدات استجابة مناعية, وذلك بالتفاعل مع هذه المستقبلات، ويعرف هذا التفاعل بأنه التحفيز [7]. يطلق على الأجزاء الصغيرة على سطح المُستضد بالحَواتِم (موقع الاستضدادية في الجزيئة) epitopes والأجزاء الصغيرة على الأجسام المضادة يطلق عليها المُستوقِعات (موقع الارتباط بالمُستضد) paratopes، كذلك تعرف بأنها المنطقة المتغيرة V–region، وهي مسؤولة عن تطابق (تمييز) المُستضد، كما أنها قابلة للتغير؛ لأنها من الممكن أن تغير شكلها للحصول على أفضل تطابق (تتميم) مع المُستضد المعطى كما مبين في الشكل (1) إذ تطلق الحَواتِم استجابة مناعة معينة و المُستوقِعات للأجسام المضادة مرتبطة بهذه الحَواتِم بموجب قوة رابطة معينة، تعرف بأنها درجة الصلة affinity, تقاس قوة وخصوصية تفاعلAb-Ag عن طريق قياس درجة صلة تطابقهما.[13،23،29].



الشكل (1) المُعقد الجسم المضاد- المُستضد[6]

إن الخلايا البائية مسؤولة عن إنتاج الأجسام المضادة وإفرازها وهي بروتينات خاصة ترتبط بالمُستضد, ويمكن لكل خلية بائية أن تنتج جسماً مضاداً خاصاً واحداً. تتضمن النظرية التقليدية آلية الانتقاء السلبي والإيجابي.

1.1.4. آلية الانتقاء السلبي

إن هدف الانتقاء السلبي هو تجهيز قدرة تحمل(ميزة السماحية) للخلايا الذاتية؛ إذ تتعامل مع قدرة جهاز المناعة على اكتشاف المُستضدات غير المعروفة في حين لا يتفاعل مع الخلايا الذاتية في الوقت نفسه [9]؛ وأثناء تولد الخلايا التائية تنشأ المستقبلات من خلال عملية إعادة تنظيم جينية عشوائية. وبعد ذلك تقوم بإجراء عملية مراقبة من التوتة (الغدة الصعترية) تدعى الانتقاء السلبي؛ إذ أن الخلايا التائية التي تتفاعل ضد البروتينات الذاتية لا يُسمح لها بمغادرة التوتة وتدمر بينما الخلايا التائية التي لا ترتبط بالبروتينات الذاتية هي فقط التي يُسمح لها بمغادرة التوتة. ثم تنتشر هذه الخلايا التائية الناضجة في جميع أنحاء الجسم لتؤدي وظائف مناعية وتحمي الجسم من المُستضدات [1، 26, 28]. كذلك يمكن للخلايا البائية أن تصبح قادرة على التحمل إذا ما واجهت المُستضد في غياب كل من الخلية التائية المساعدة والتأثيرات التحفيزية. وكما هو الحال مع الخلايا التائية، فإنه يُمكن للخلايا

البائية المتفاعلة ذاتياً أن تفر من الانتقاء السلبي المركزي للخلية البائية, وفي هذه الحالة ستكون فعالية الخلية البائية أو قدرة تحملها هي نتيجة العدد والقوة والوقت الذي ستظهر عنده إشارات تحفيزية مساعدة.

2.1.4. آلية الانتقاء الايجابي

يهدف الانتقاء الايجابي للخلايا اللمفاوية (البائية والتائية) إلى تجنب تجميع الخلايا اللمفاوية غير المفيدة والتي إما لا تمتلك مُستقبِلاً مطلقاً أو أن مُستقبِلاتها غير منتجة للكائن الحي ونتيجة لذلك فإن الخلايا التي تَنْجو من الانتقاء الايجابي ستنجو من موت الخلية وتصبح أكثر فعالية في عملية تمييز المُستضد [6]. يجب أن تميز جميع خلايا التائية المُستضدات المرتبطة مع جزيئات MHC الذاتي الذي يدعى بمعقدات البيتيد / MHC – الذاتي؛ لذلك فمن الضروري اختيار الخلايا التائية غير الناضجة، التي تسمى الخلايا التوتية thymocytes، والتي تكون مستقبلاتها قادرة على التمييز والارتباط مع جزيئات MHC – الذاتية [24]. تضمن عملية الانتقاء الايجابي أن الخلايا التائية الناضجة التي ستغادر التوتة (الغدة الصعترية) وتنتشر في جميع أنحاء الجسم سوف تَتَفْعّل فقط بواسطة المُستضدات الغريبة التي تقدمها جزيئات MHC – الذاتية, يتضمن أيضاً الانتقاء الايجابي الخلايا البائية الناضجة وإنقاذها من الموت.

عند المقارنة مع الخلايا التائية، فإن الانتقاء الايجابي للخلايا البائية الناضجة يبدو مماثلاً جداً للانتقاء التوتي الايجابي للخلايا التائية غير الناضجة. ففي حالة الخلية التائية، تُنقذ الخلايا اللمفاوية من موت الخلية بسبب تمييزها لجزيئة MHC – الذاتية، بينما في حالة الخلية البائية، تُنقذ الخلايا البائية من الموت بسبب تمييزها الجزيئات غير الذاتية بوجود الإشارات التحفيزية [6].

2.4. نظرية الانتقاء النَسيليّ

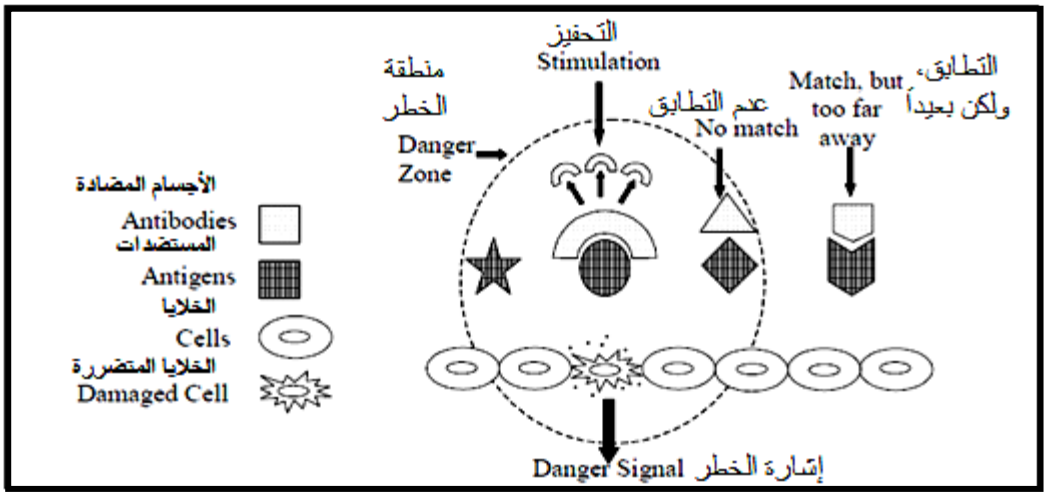
يصف مبدأ الانتقاء النَسيليّ الصفات الأساسية لاستجابة المناعة لمحفزات المُستضد، الفكرة الأساسية لهذه النظرية أنه يتم انتقاء وتكاثر الخلايا اللمفاوية القادرة على تمييز المُستضدات بدلاً من الخلايا التي لا تقوم بذلك وتتمايز إلى الخلايا المُستَفعِلة Effector Cells [16].

3.4. نظرية شبكة المناعة

تعد نظرية الانتقاء النَسيليّ لجهاز المناعة مجموعة من الخلايا والجزيئات المنفصلة وهي بالأصل في استراحة ويتم إثارتها فقط بوساطة تحفيز مستضد غريب, تقدم شبكة المناعة نظرية مختلفة فكرياً لكيفية تفاعل مكونات جهاز المناعة مع بعضها البعض ومع المحيط (المُستضدات) [6], لقد اقترح Jerne [30] أنموذج شبكة المناعة, وبَيّنَ أن جهاز المناعة يحتفظ بشبكة نوعية مميزة من الخلايا المترابطة لتمييز المُستضد. تتحفز هذه الخلايا وتثبط إحداها الأُخرى بطريقة معينة تؤدي إلى استقرار الشبكة.

4.4. نظرية الخطر

قدمَت عالمة المناعة Polly Matzinger في عام 1994 نظرية الخطر [22]. ذكرت أن جهاز المناعة يُسَيطَر عليه بوساطة اكتشاف الضرر الذي يحصل في الجسم، وليس بوساطة اكتشاف تراكيب المُستضد وأن الإشارات لا تأتي من مصادر خارجية المنشأ، ولكنها داخلية وتقوم بانتاجها خلايا النسيج نفسها. إن هذه الإشارات الداخلية تسمى إشارات الخطر danger signals [12].يوضح الشكل (2) نظرية الخطر [27], وتعمل الخلية التغصنية بمبدأ نظرية الخطر.

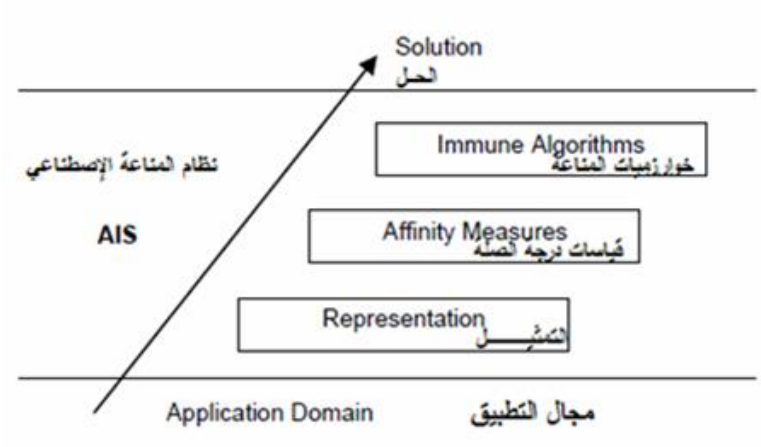


الشكل (2) أنموذج نظرية الخطر[27]

5. نظام المناعة الاصطناعي

AIS هو منهج ذكي متطور حديث، مستوحى من جهاز المناعة الطبيعي في الإنسان، ظهر في التسعينات على أنه فرع جديد من الحساب التطوري Evolutionary Computation[40]. إن جهاز المناعة الطبيعي معقد جداً ليتم مُحاكاته صناعياً، ولكن نجح- A.B. Watkins بمحاكاة أهم وظائف جهاز المناعة الطبيعي فيما يخص تمييز الأنماط. إن نظام المناعة الاصطناعي، هو مجموعة من خوارزميات التصنيف الذكية، التي تستخدم آلية دفاع المناعة الطبيعية لأغراض تقانية، قادرة على التكيف والتعلم. لذلك انتشر هذا المفهوم في تطبيقات تقانة عديدة في العقود الأخيرة [19]. إن الحوسبة المستوحاة إحيائياً (Biologically Inspired Computing)، وخصوصاً أنظمة المناعة الاصطناعية (AIS) هي حل واعد لتطوير أدوات دفاعية مكيفة ومؤتمتة للتهديدات الحالية والمستقبلية في أكبر أنظمة (تقانة المعلومات)[33] .

قدم De.Castro وTimmis [6] إطار عمل مستخدم بصورة أكثر شيوعاً في هندسة AIS، والموضح في الشكل (3).



الشكل (3) طبقات إطار عمل AIS [6]

يتضمن إطار العمل ثلاث خطوات مستقلة بشكل نسبي في بناء (AIS). إن أساس كل نظام هو مجال التطبيق، يعمل تمثيل البيانات على التعريف بكيفية تحول بيانات التطبيق الملاحظة إلى بيانات متلائمة النسق

وذلك لتعالج بوساطة خوارزميات (AIS). تحتاج البيانات المستخدمة أن تكون مُنسقة لتكون قادرة على قياس درجة الصلة في مجال تمثيل البيانات، إن قياس درجة الصلة هو في صلب التصنيف والتمييز بوساطة خوارزميات (AIS) (على الأقل هذا صحيح بالنسبة لخوارزميات (AIS) التكيفية) [6].

6. مصنف أنموذج الانتقاء السلبي والايجابي الهجين

طور في هذا البحث أنموذج التصنيف HPNS الذي يعتمد على نظام المناعة التكيفي؛ إذ تم في هذا الأنموذج التهجين بين خوارزمية الانتقاء السلبي (NSA) وخوارزمية الانتقاء الايجابي (PSA) لحل مشكلة تصنيف الوثائق وهما إحدى خوارزميات النظام المناعي التكيفي؛ إذ إن آلية الانتقاء السلبي هي عملية انتقاء الخلايا التائية التي لا تطابق مستقبلات الخلايا التائية في المخزن، في حين أن آلية الانتقاء الايجابي هي عملية اختيار الخلايا التائية التي تطابق مستقبلات الخلايا التائية الموجودة في المخزن، وكلا الآليتين تحدث في الغدة التوتية.

لقد وظفت الفكرة ذاتها؛ إذ يقوم أنموذج النظام المناعي المصنف المبني على الخوارزمية الهجينة(HPNS)، والمصمم في هذا البحث بتصنيف وثائق مرحلتي هندسة البرمجيات وهما المتطلبات(SRS) (Software Requirement Specification) والاختبار (STD) (Software Test Document) .

يوضح الجدول (1) عملية التماثل بين نظام التصنيف المناعي الاصطناعي والجهاز المناعي الطبيعي. والشكل (4) يوضح تحليل عمليات أنموذج HPNS.

الجدول (1) التماثل بين جهاز المناعة الطبيعي و نظام التصنيف المناعي الاصطناعي

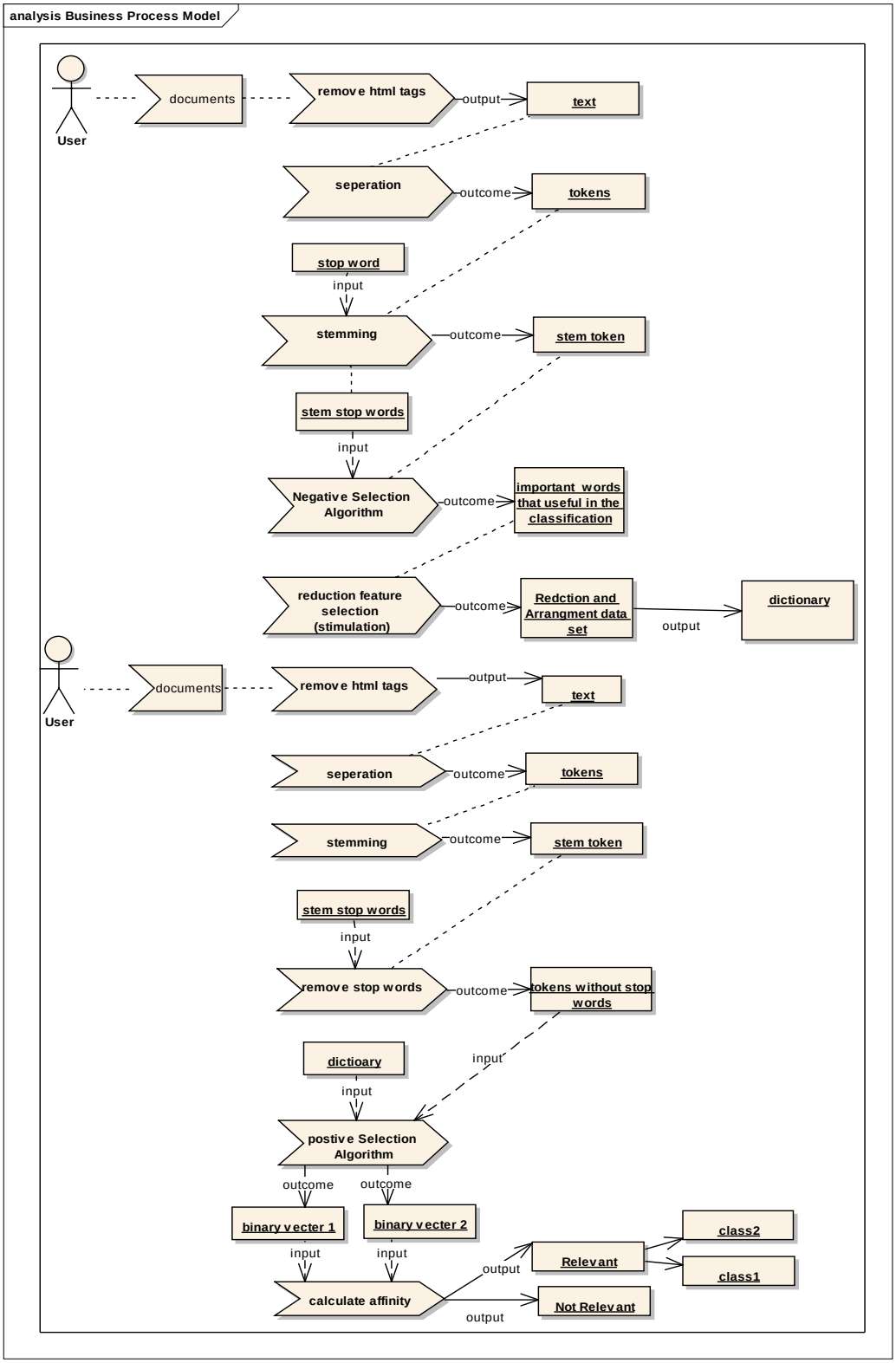
العناصر	نظام المناعة الطبيعي	نظام المناعة الاصطناعي
1- الإدخال	المستضدات	الوثائق
2- المعلومات المخزونة	الخلايا التائية الذاتية (غير ناضجة) الموجودة في مخزن الانتقاء السلبي	Stop words
3- البيانات المدربة	الخلايا التائية الناضجة الموجودة في مخزن الانتقاء الايجابي	القاموس
4- البيئة	الغدة التوتية	نظام التصنيف
5- أسلوب الرد	استجابة نظام المناعي الطبيعي	ناتج نظام التصنيف
6- جزء من بيانات الإدخال	الببتيدات	متجه الخواص

1.6. وصف أنموذج الانتقاء السلبي والايجابي الهجين

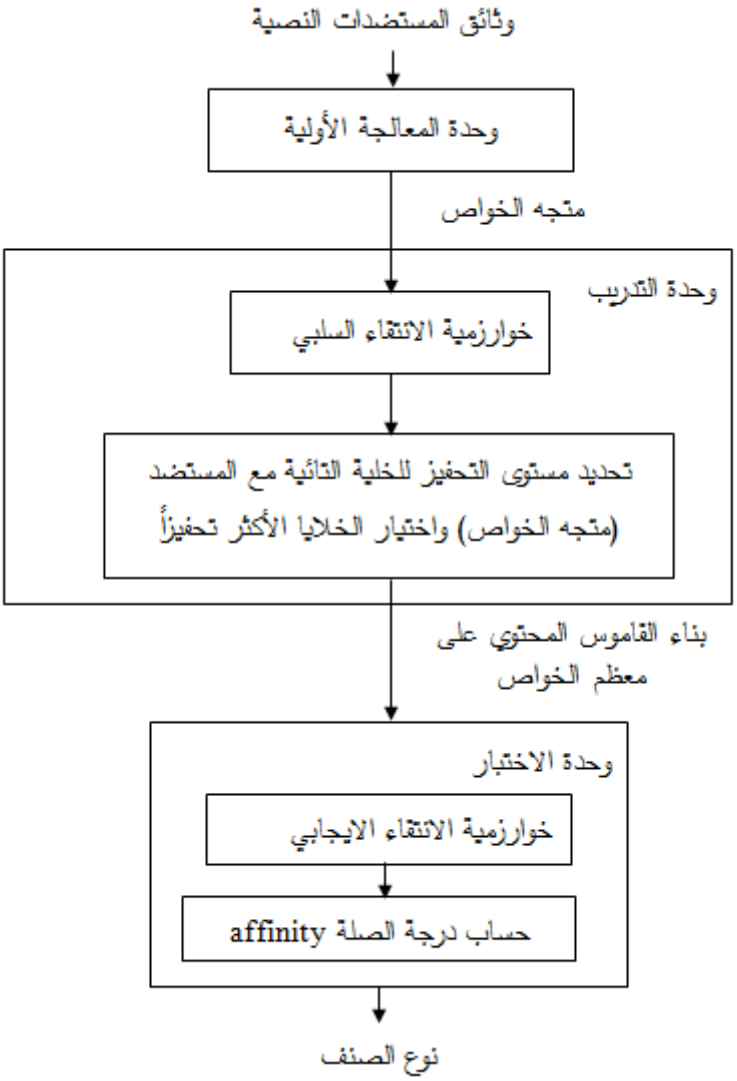
تتكون خوارزمية التصنيف التكيفي (HPNS) من ثلاث وحدات أساسية:

1- وحدة المعالجة الأولية.
2- وحدة التدريب .
3- وحدة الاختبار.

تتفاعل هذه الوحدات مع بعضها البعض لانجاز مهمة تصنيف الوثائق والشكل (5) يوضح عمل الوحدات بصورة عامة.



الشكل(4) أنموذج عمليات HPNS



الشكل (5) التفاعل بين وحدات HPNS

* **<u>وحدة المعالجة الأولية</u>** : تدخل الوثيقة النصية المراد تصنيفها إلى وحدة المعالجة الأولية وذلك لتهيئتها وتحويلها إلى صيغة يسهل على وحدة التدريب التعامل معها, إن معالجة الوثيقة النصية هي خطوة مهمة جداً في التصنيف ، والهدف الرئيس منها هو إزالة الصفات (الكلمات) التي لا تعطي أي معلومات عن صنف الوثيقة, فضلاً عن إزالة البيانات المتكررة, تتضمن عملية المعالجة استخلاص الخواص عن طريق تحويل وثيقة النص المدخلة إلى متجه خواص الصفات، ويضم التحويل الخطوات التالية:

1- تحويل الوثيقة من نوع (HTML) إلى نص واضح (Plain Text) وذلك بإزالة الـ(Java Script) والـ(HTML tags) .
2- إزالة الأرقام وعلامات التنقيط وتقسيم النص إلى كلمات منفصلة.
3- ازالة التكرار وتكوين قائمة بكلمات الوثيقة.
4- تحويل الكلمات الموجودة في القائمة إلى أحرف كبيرة (Upper Case Letters)
5- تطبيق خوارزمية الـ(Stemming) لإعادة الكلمات إلى جذرها مثل Testing, Tested, Testes → Test

* **وحدة التدريب** : تتألف من مرحلتين:

المرحلة الأولى : استخدام خوارزمية الانتقاء السلبي (NSA) لتوليد القاموس (الكاشف) لكل صنف.

المرحلة الثانية : حساب مستوى التحفيز للخلايا التائية (كلمات القاموس) واختيار الخلايا ذات التحفيز العالي (الكلمات المهمة في التصنيف).

في المرحلة الأولى تُستخدم الخلايا التائية الذاتية (Stop words) لفحص عينات الإدخال الجديدة (المستضد ≡ متجه الخواص)، إذا طابقت الخلايا التائية المستضد (كلمة من الـStop words طابقت كلمة في متجه الخواص) عندها تزال تلك الكلمة (ترفض هذه الخاصية) وإلا تقبل (إضافة الخاصية إلى القاموس بوصفها كاشفاً جديداً)، وهذا يقابل خاصية النظام المناعي الطبيعي ذاتي/غير ذاتي، والمخطط الانسيابي (6) يوضح خوارزمية الـ (NSA).

وفي المرحلة الثانية حُسب مستوى التحفيز للخلايا التائية (كلمات القاموس) واختيار الخلايا ذات التحفيز العالي (الكلمات المهمة في التصنيف) باستخدام طريقة Mutual Information (MI) كما مبين في المعادلة رقم(1)[25].

$$MI(t,c) = \log \frac{P(c,t)}{P(t).P(c)} \quad .....(1)$$

إذ أن:

t : الكلمة

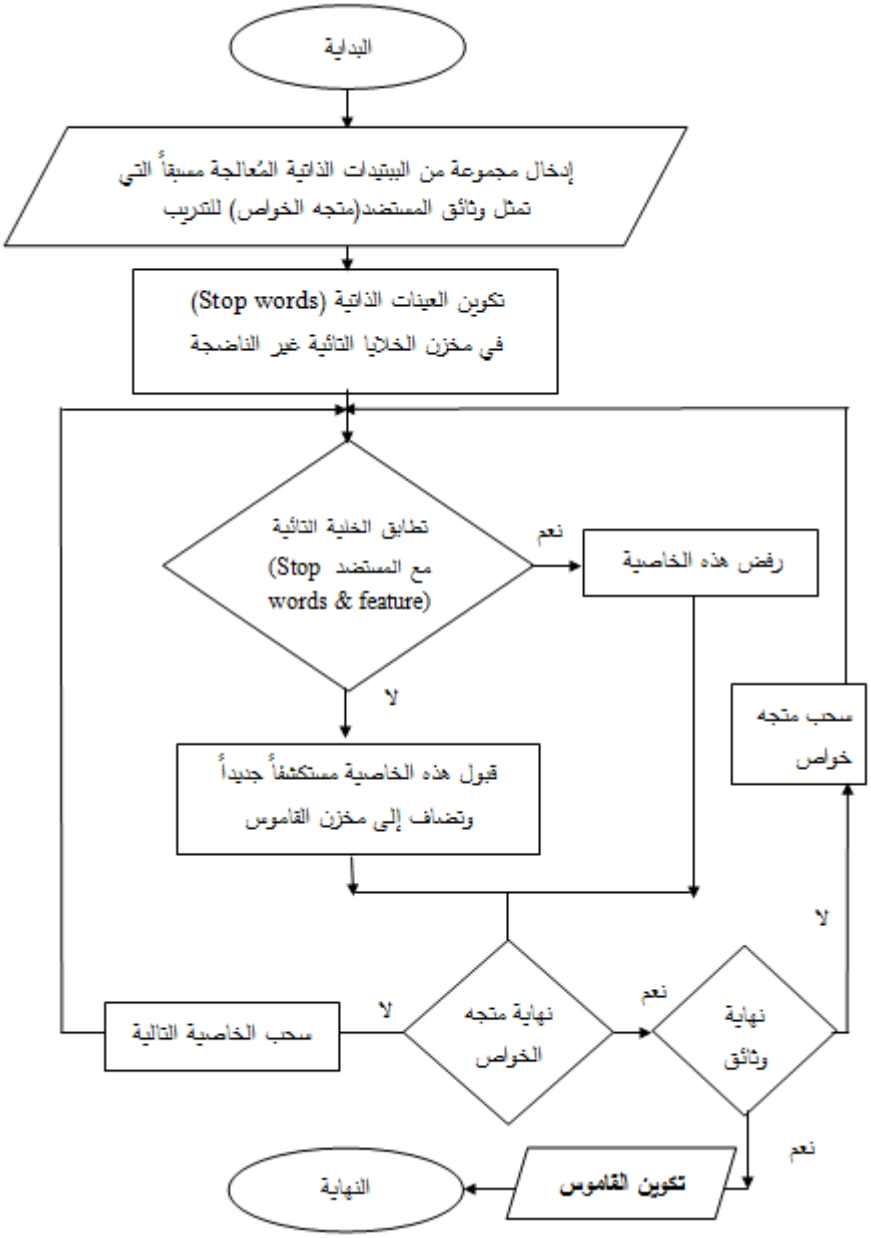
c: الصنف

P(t,c): احتمالية وجود الكلمة في الصنف c = عدد الوثائق التي تحوي الكلمة t بالصنف c / عدد الوثائق التي تعود للصنف c

P(t): احتمالية وجود الكلمة في كل الوثائق = عدد مرات تواجد الكلمة في الصنفين / عدد الوثائق الكلي

P(c): احتمالية عدد الوثائق التي تعود للصنف c = عدد وثائق الصنف c / عدد الوثائق الكلي

بعد تطبيق طريقة (MI) على كواشف القاموس أصبح لكل كاشف (كلمة) قيمة تحفيز خاصة بها, وبالاعتماد على هذه القيم ترتب الكلمات تنازلياً وأخذ أهم الكلمات حسب الاختبارات التي أُجريت على البيانات المستخدمة في البحث والتي تمتلك أكبر قيمة من المعلومات التصنيفية (أعلى مستوى تحفيز) وإهمال البقية؛ إذ كلما زاد مستوى التحفيز زاد مستوى أهمية الكلمة في التصنيف ، وبهذا يتقلل حجم القاموس الذي يؤثر في دقة التصنيف وسرعته بصورة إيجابية، وتمثل هذه الكواشف (الكلمات) الخلايا التائية الناضجة في النظام المناعي.



الشكل (6) المخطط الانسيابي لخوارزمية الـ(NSA)

بعد تطبيق طريقة (MI) على كواشف القاموس أصبح لكل كاشف (كلمة) قيمة تحفيز خاصة بها وبالاعتماد على هذه القيم ترتب الكلمات تنازلياً وتؤخذ أهم الكلمات حسب الاختبارات التي أُجريت على البيانات المستخدمة في البحث والتي تمتلك اكبر قيمة من المعلومات التصنيفية (أعلى مستوى تحفيز) وإهمال البقية، إذ كلما زاد مستوى التحفيز زاد مستوى أهمية الكلمة في التصنيف، وبهذا يقلل حجم القاموس الذي يؤثر في دقة التصنيف وسرعته بصورة إيجابية، وتمثل هذه الكواشف (الكلمات) الخلايا التائية الناضجة في النظام المناعي.

* وحدة الاختبار : تتضمن استخدام خوارزمية الانتقاء الايجابي (PSA) وتحديد مقدار الصلة بين المستضد(وثائق الاختبار) والخلايا التائية لتحديد الصنف الذي تعود إليه الوثيقة, ويعمل الانتقاء الايجابي على تحفيز الخلايا التائية الناضجة القادرة على تمييز المستضد. إذا لم تميز الخلية التائية مستضداً واحداً على الأقل يتم

حذفها وإلا يتم إختيارها بوصفها خلايا مؤهلة مناعياً Immune Competent Cell ثم حساب قيمة ترابط المستضد مع الخلايا التائية والمخطط الانسيابي(7) يوضح خوارزمية الانتقاء الايجابي مع حساب قيمة التطابق.

تُحول خوارزمية الانتقاء الايجابي متجه خواص الكلمات النصية إلى متجه ثنائي بالاعتماد على كواشف القاموس، فإن وجدت الخاصية في القاموس تمثل بالقيمة (1) في المتجه الثنائي وإذا لم توجد يتم تمثيلها بـ(0).

تحسب درجة الصلة (Affinity) للمتجه الثنائي مع الخلايا التائية صنف1 (القاموس1) ومع الخلايا التائية صنف2 (القاموس2) إذ إن المتجه ذا الصلة الأعلى تعني أن الوثيقة تعود لذلك الصنف، وذلك باستخدام قانون البتات المستمرة المتعددة:

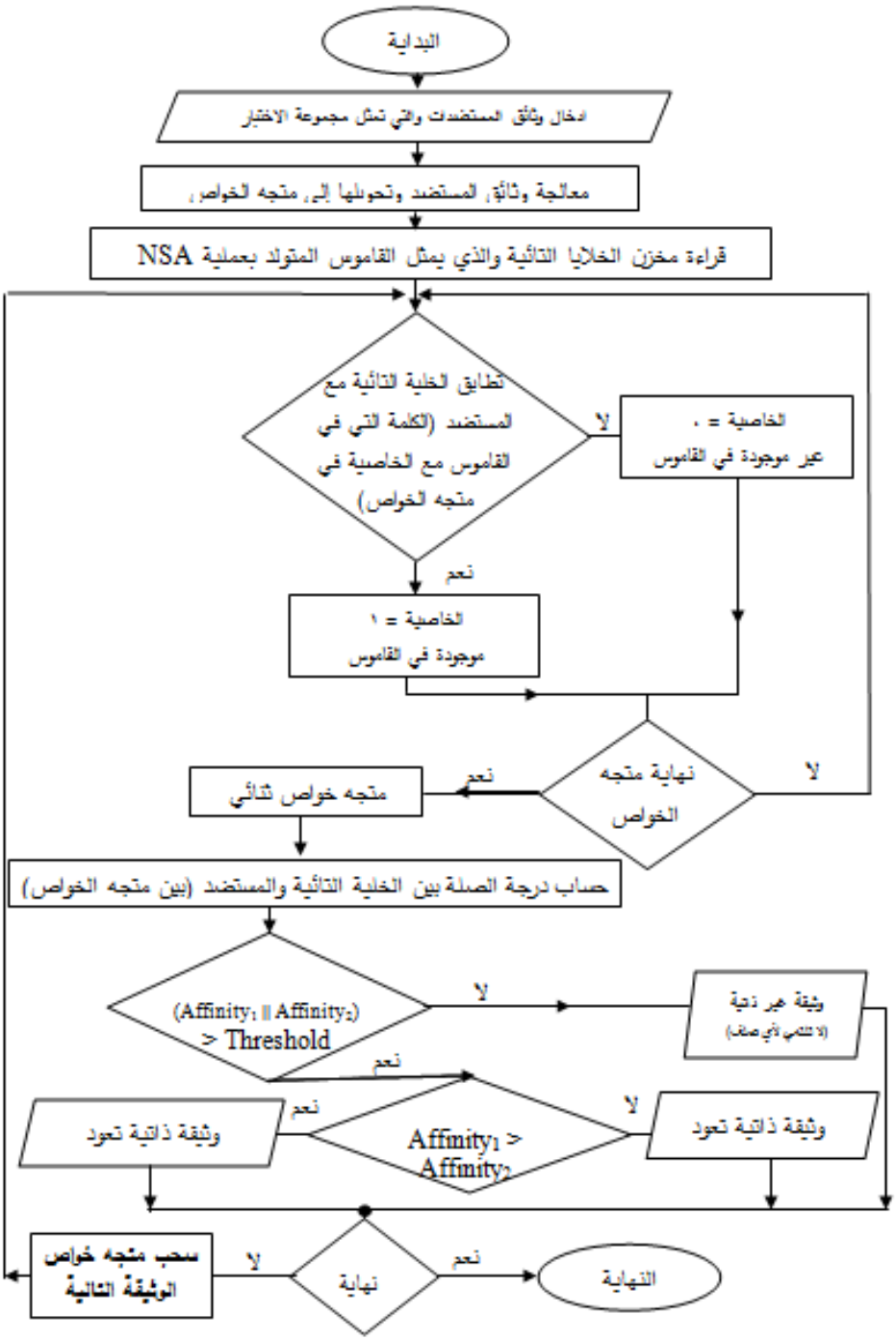
قانون البتات المستمرة المتعددة = عدد الواحدات الكلي + 2عدد الواحدات المستمرة .....(2)

بعد إختيار الخوارزمية المناسبة لكل وحدة وتمثيل العناصر يتم البدء بتحديد تفاصيل التصميم؛ إذ يتكون الأنموذج الأول للصنف من ثلاثة حزم (Package) :

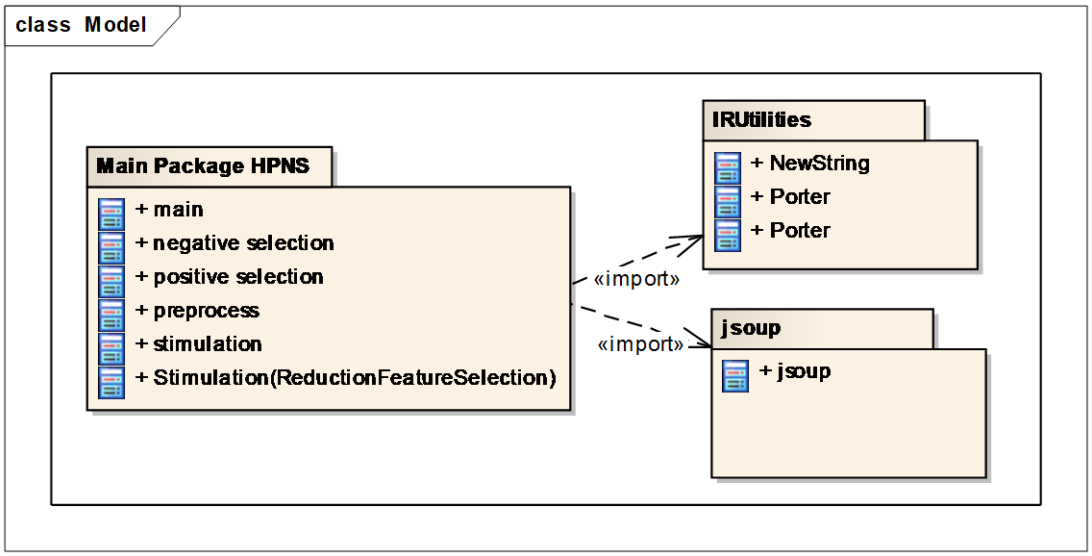
- **الحزمة الأولى** : هي الحزمة الرئيسة التي تحوي استدعاء كل صنفيات (Classes) المعالجة الأولية ، التدريب ، التحفيز ، الاختبار واستدعاء الحزم الأخرى .
- **الحزمة الثانية** : هي حزمة IRutilities التي تحوي خوارزمية Porter الخاصة بعملية (Stemming) .
- **الحزمة الثالثة** : هي حزمة Jsoup التي تقوم بتحويل وثيقة html إلى سلسلة (String) .

والشكل (8) يوضح حزم أنموذج HPNS المستخدمة في تصنيف الوثائق والصنفيات (Classes) داخل كل حزمة .

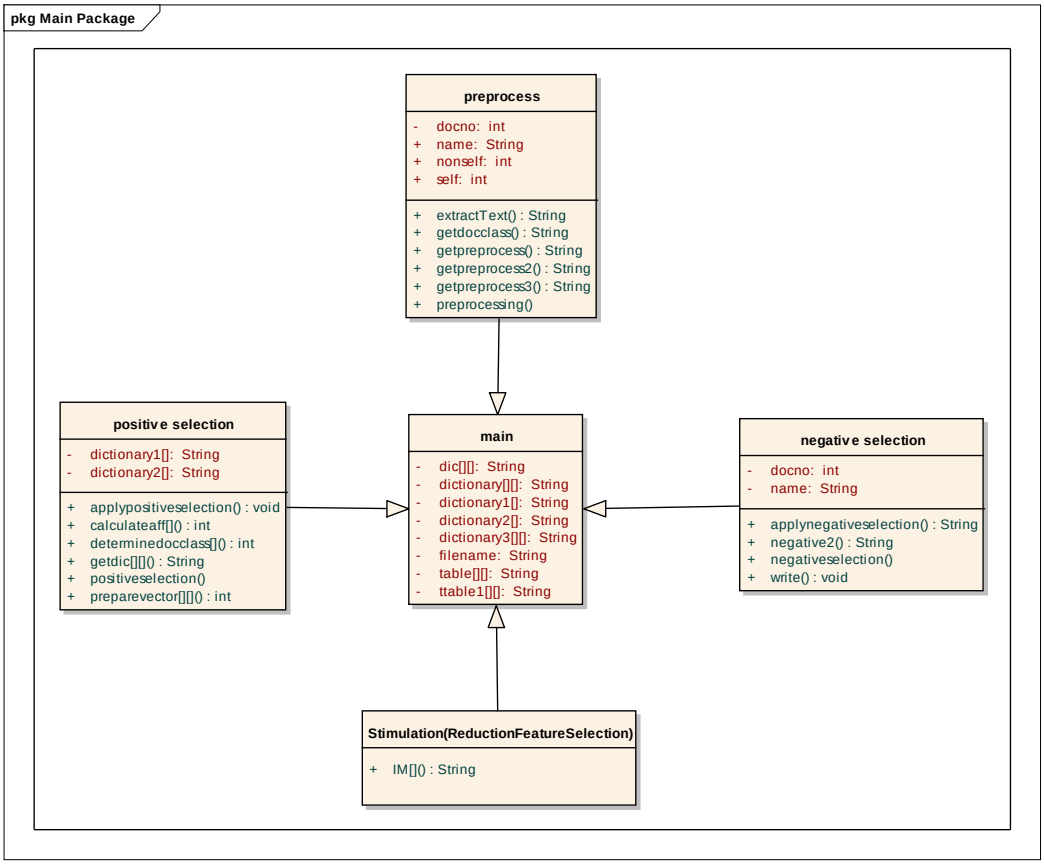
بعدها تحدد المتغيرات والدوال الخاصة بكل صنف (Class) باستخدام مخطط الصنفيات الموضح في الأشكال (9)، (10)، (11) .



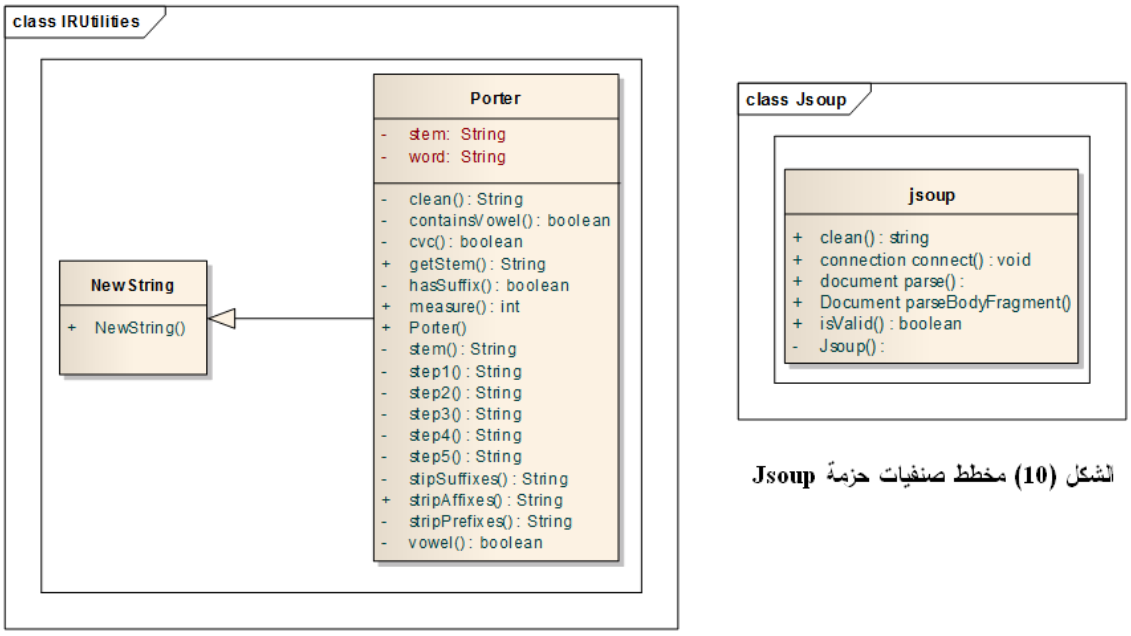
الشكل (7) المخطط الانسيابي لخوارزمية (PSA) وحساب مقدار صلة المستضد والخلية التائية



الشكل (8) مخطط الصنفيات للحزم الثلاثة لأنموذج HPNS



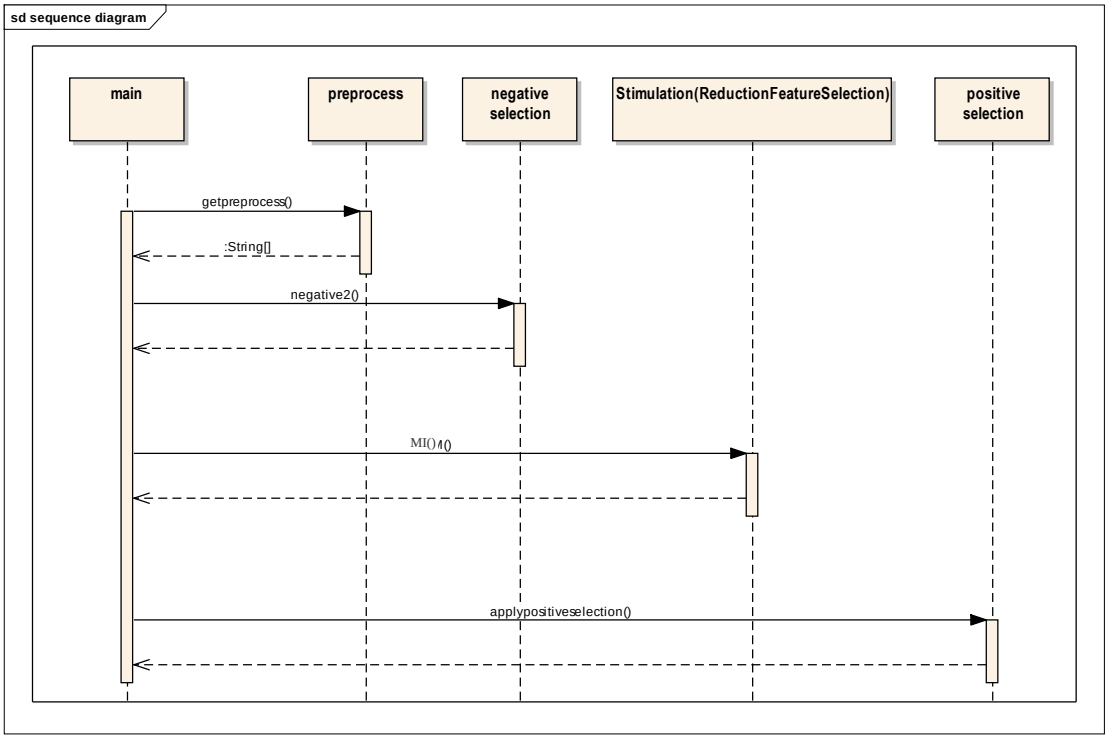
الشكل (9) مخطط صنفيات الحزمة الرئيسة لأنموذج HPNS



الشكل (10) مخطط صنفيات حزمة Jsoup

الشكل (11) مخطط صنفيات حزمة IRUtilities

ثم يحدد الترتيب الزمني للصنفيات حسب التسلسل الزمني لاستدعائهم باستخدام المخطط التسلسلي الموضح في الشكل (12) .



الشكل (12) المخطط التسلسلي لأنموذج HPNS

* **مرحلة بناء HPNS :** تضم هذه المرحلة كتابة البرنامج ويتم فيها تحويل الخوارزميات والمخططات المحددة في مرحلة البناء للطور الأول إلى مقاطع برمجية (Source Code) وذلك بإستخدام لغة Javasun .

* **مرحلـــــــة الاختبـــــــار**: يختبـــــر البرنـــــامج باســـــتخدام اختبـــــار الصـــــندوق الأســـــود (Black Box Test) ؛ الذي لا يهتم بما في داخل البرنامج ولا كيف صمم وأنشئ ولكنه يهتم فقط بوظائف البرنامج هل يقوم بها على أكمل وجه أم لا، فضلاً عن اختبار أداء الأنموذج، وهو تقييم البرنامج بوصفه نظاماً يقوم بوظائف محددة حسب المتطلبات ويندرج هذا الاختبار باختبار الوحدات الثلاثة :

- **المعالجة الأولية** :- الإدخال وثيقة نص بصيغة html، الاختبار هو التأكد من أن الإخراج يكون قائمة من الكلمات (متجه الخواص) بحالة الأحرف العلوية خالية من علامات التنقيط ، الأرقام والتكرار . فضلاً عن إرجاع الكلمة إلى أصلها (جذرها) .
- **التدريب** : الإدخال قائمة الكلمات، الاختبار هو التأكد من بناء قاموس خالٍ من الكلمات التي ليس لها أهمية في التصنيف (Stopwords) وتحوي فقط الكلمات ذات الأهمية الأكبر.
- **الاختبـار** : الإدخال وثائق نصية بصيغة html ، الاختبار التأكد من أن البرنامج يقوم بتصنيف الوثائق لصنفها الصحيح بالاعتماد على درجة الصلة بينها وبين القاموس المكون في مرحلة التدريب، وتقييم البرنامج.

يُقيم أداء الطريقة باستخدام المعايير الموضحة في الفقرة التالية؛ ومن هذه المقاييس الدقة و Recall, F_1-macro, F_1-micro, Specificity و(Tp, Tn, Fn, Fp).

وفي حالة كشف أي خطأ في البرنامج يتم الرجوع للخطوة السابقة ومعالجة المشكلة ثم التأكد مرة اخرى من صحته.

*** مقاييس التقييم**

هنالك طرائق عديدة لتحديد الفعالية إلا أن المقاييس الأكثر استخداما هي (accuracy، recall ، precision). ولأجل حساب هذه المقاييس يجب تحديد قيم True Positive(TP) ايجابي صحيح او False Positive (FP) ايجابي خاطئ أو True Negative (TN) سلبي صحيح أو False Negative (FN) سلبي خاطئ الموضح في الجدول (2)[60].

جدول (2) تصنيف الوثائق [60]

الرمز	التوضيح
TP	عدد الوثائق التي تم تصنيفها بشكل صحيح
FP	عدد الوثائق التي تم تصنيفها بشكل غير صحيح
FN	عدد الوثائق التي يجب أن تعود لصنف معين لكنها لم تصنف كذلك
TN	عدد الوثائق التي يجب أن لا تعود لصنف معين وهي كذلك

.

TP : عدد الوثائق التي تم تصنيفها بشكل صحيح.

FP : عدد الوثائق التي تم تصنيفها بشكل غير صحيح.

FN : عدد الوثائق التي يجب أن تعود لصنف معين لكنها لم تصنف كذلك.

TN : عدد الوثائق التي يجب أن لا تعود لصنف معين وهي كذلك.

Precision : يحـــــدد علـــــى انهـــــا النســـــبة التـــــي يـــــتم تصـــــنيف وثيقـــــة عشـــــوائية dx وفقاً لـ ci، أو ما يمكن ان يعد صنفاً صحيحاً.

$$\pi_i = \frac{Tp_i}{Tp_i + FP_i} \quad ....(3)$$

Recall : تعرف على انها النسبة التي يتم فيها اتخاذ قرار اذا كان يجب تصنيف وثيقة عشوائية dx تحت صنف (ci).

$$P_i = \frac{Tp_i}{Tp_i + FN_i} .....(4)$$

Accuracy : غالباً ما يستخدم على اعتبار انها مقياس لأساليب التصنيف الا ان قيم Accuracy اكثر تقبلاً للتباينات في عدد القرارات الصحيحة مقارنة بـ Precision و Recall.

$$A_i = \frac{Tp_i + TN_i}{Tp_i + TN_i + Fp_i + FN_i} .....(5)$$

فضلا عن ذلك يتم دمج π_i و P_i في مقياس واحد لكي يعطي صورة افضل عن اداء المصنف ويسمى مقياس F1 والذي يتم حسابه بطريقتين Micro-average و Macro-average [30]:

1- Micro-average F1-measure : يقوم بحساب قرار التصنيف بشكل عام للصنفين كما في المعادلات(6) الى (8).

$$\text{F1(micro-averaged)} = \frac{2\pi p}{\pi + p} .....(6)$$

$$\pi = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\sum_{i=1}^{M} \text{TP}_i}{\sum_{i=1}^{M} (\text{TP}_i + \text{FP}_i)} ....(7)$$

$$\text{P} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\sum_{i=1}^{M} \text{TP}_i}{\sum_{i=1}^{M} (\text{TP}_i + \text{FN}_i)} .....(8)$$

2- Macro-average F1-measure : يقوم بحساب قرار التصنيف بالنسبة لكل صنف ثم يتم اخذ المعدل لهما كما في المعادلات(9) الى (10).

$$F1_i = \frac{2\pi_i P_i}{\pi_i + P_i} .....(9)$$

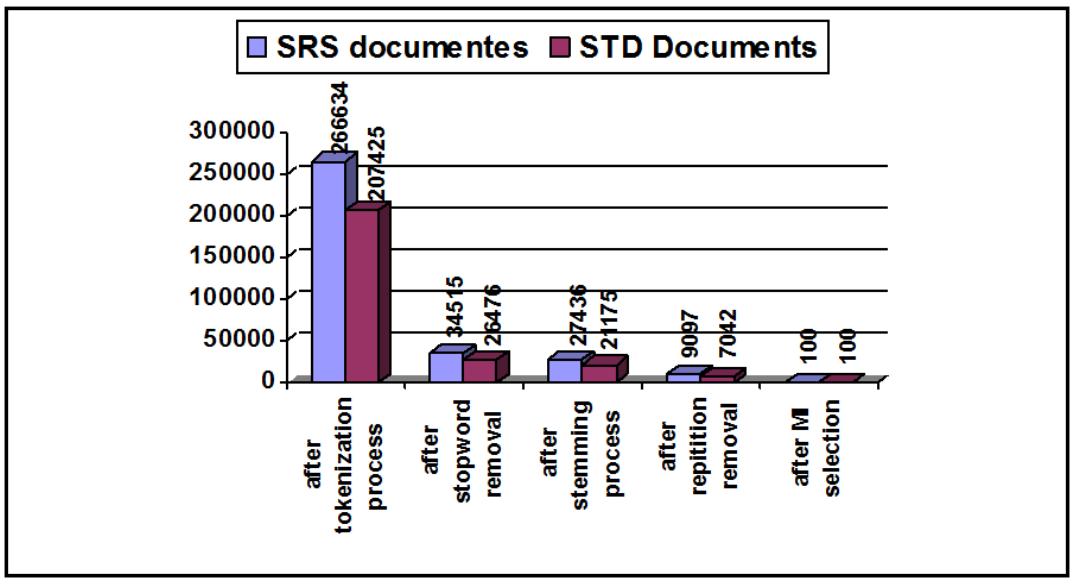
$$F1 \text{ (macro – averaged)} = = \frac{\sum_{i=1}^{M} F1_i}{M} ....(10)$$

M = عدد الاصناف

2.6. تهيئة وتنفيذ أنموذج HPNS Initialization & Implementation of HPNS Model

هيئ ونفذ الأنموذج باستخدام 80 وثيقة: 32 وثيقة من صنف Requirements Specification (SRS) Software و 48 وثيقة من صنف Software Test Description (STD).

يوجد صيغ عديدة لخزن الوثائق ومن أشهرها (HTML)، وعند تهيئة النظام باستخدام هذه الوثائق تمرر بعدة مراحل لتحويلها إلى صيغة يسهل على المصنف التعامل معها. والعديد من الخواص المهمة يتم الحصول عليها خلال مرحلة التهيئة، والخطوة الأساسية في هذه المرحلة هي تحديد الخواص التي ستستخدم في مرحلة التصنيف. والحل الأبسط هو استخدام كل الخواص المتمثلة في مجموعة وثائق التدريب، لكن إذا كان عدد الخواص كبيراً جداً يقلل باستخدام تقانات مختلفة بخطوات متعددة وهي إزالة كلمات stop word، إعادة الكلمات إلى جذرها stemming واختيار الخواص التي تمتلك أعلى معدل من المعلوماتية MI التي تفيد في التصنيف وخزنها في القاموس، والشكل(13) يوضح تأثير كل من هذه التقانات على تقليل عدد الخواص.



الشكل (13) تأثير العمليات المختلفة في حجم متجه الخواص

3.6. اختبار الأنموذج ونتائجه Testing of Model and Results

اختبر الأنموذج HPNS بإدخال مجموعة من وثائق مراحل هندسة البرمجيات التي حملت من المواقع الالكترونية الخاصة بوثائق هندسة البرمجيات، وتصنف الوثائق المختارة إلى واحد من الأصناف التالية :

1- الصنف الأول Software Requirements Specification (SRS).

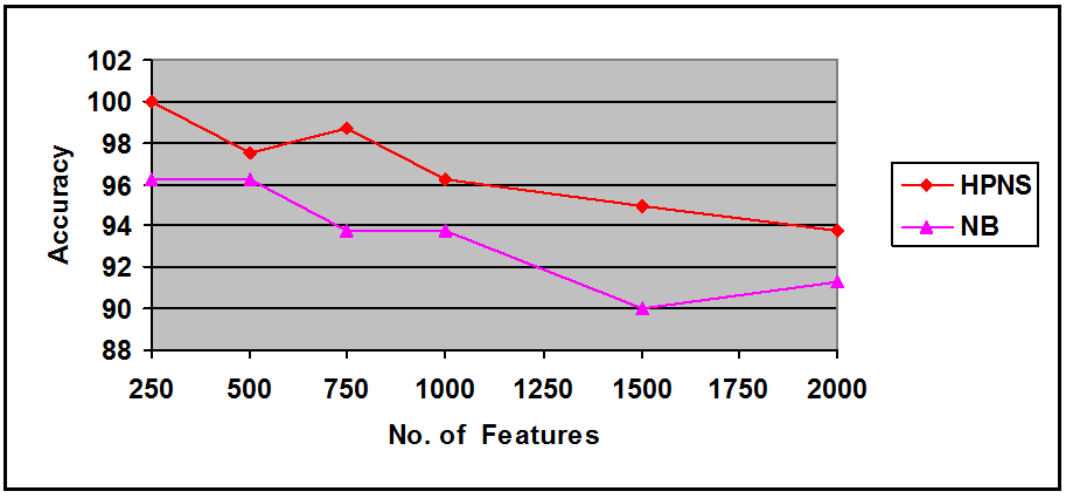
2- الصنف الثاني Software Test Description (STD) .

3- لا تعود للصنفين.

أجري الاختبار على مجاميع باعداد مختلفة من الوثائق، الجداول والأشكال التالية توضح نتائج اختبار تصنيف وثائق هندسة البرمجيات. يوضح الجدول (3) والشكل(14) نتائج دقة تصنيف وثائق التدريب من وثائق SWE لمصنفات HPNS و NB.

الجدول(3) نتائج تصنيف وثائق التدريب (80) وثيقة

Features / Model	250	500	750	1000	1500	2000
HPNS	%100	%97.5	%98.75	%96.25	%95.0	%93.75
NB	96.25	96.25	93.75	93.75	90.0	91.25

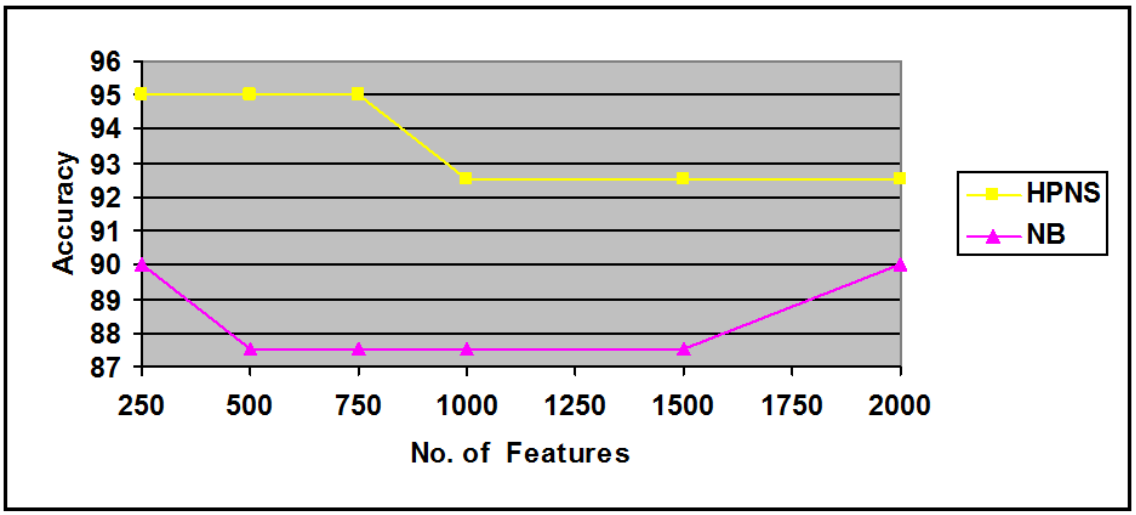


الشكل(14) دقة المصنفات لوثائق التدريب

يوضح الجدول (4) والشكل(15) نتائج دقة تصنيف (40) وثيقة اختبار من وثائق SWE للمصنفات HPNS و NB.

الجدول(4) دقة تصنيف (40) وثيقة اختبار

Features / Model	250	500	750	1000	1500	2000
HPNS	%95	%95	%95	%92.5	%92.5	%92.5
NB	90	87.5	87.5	87.5	87.5	90

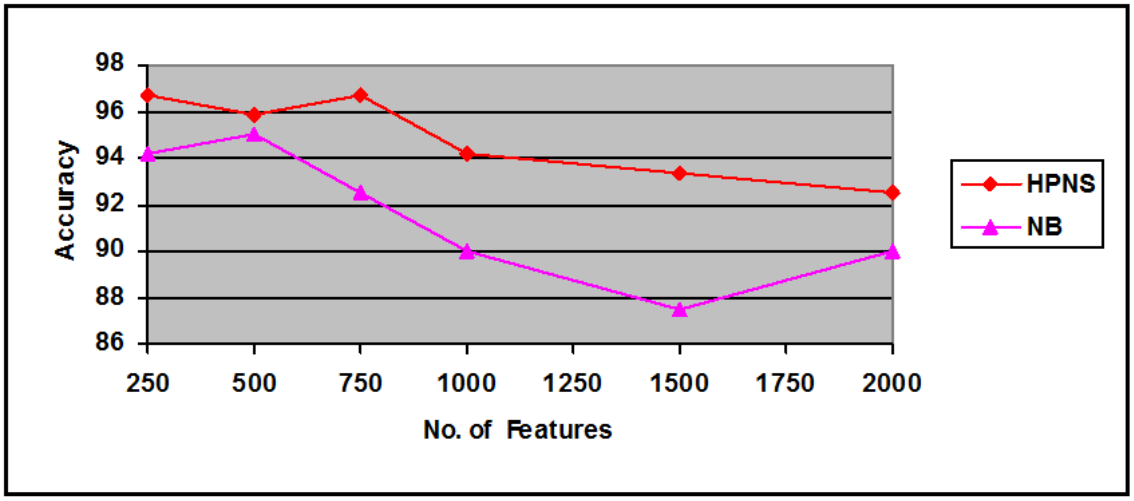


الشكل(15) دقة المصنفات لوثائق الاختبار

يوضح الجدول (5) والشكل(16) نتائج دقة التصنيف لمجاميع وثائق التدريب والاختبار الكلي والبالغ (120) وثيقة.

الجدول(5) دقة تصنيف الوثائق الكلية (التدريب والاختبار) (120) وثيقة

Features / Model	250	500	750	1000	1500	2000
HPNS	%96.67	%95.83	%96.67	%94.17	%93.33	%92.5
NB	94.17	95.0	92.5	90.0	87.5	90.0



الشكل(16) دقة المصنفات لمجموعة الوثائق الكلية

يتبين من قيم دقة المصنفات الموضحة في الجداول (2),(3),(4) أنه كلما زاد عدد الخواص كلما قلت دقة المصنف لوثائق هندسة البرمجيات، ويقدم عدد الخواص 250 و 500 أفضل دقة تنبؤ بالنسبة للمصنفات. يعد المصنف HPNS المعتمد على تقانات المناعة التكيفية أفضل من مصنف NB؛ إذ يمتلك دقة تنبؤ عالية تصل إلى 100% إذا تدرب على الوثائق مسبقاً و95% لوثائق الاختبار وتكون دقته 96.67% لمجموعة الوثائق الكلية من تدريب واختبار.

يتضمن الجدول (6) مقاييس (TP , FP , TN , FN) لوثائق الاختبار(40) وثيقة لـ (250) خاصية.

الجدول (6) مقاييس (TP , FP , TN , FN) لوثائق الاختبار

Measure / Model	TP%	FP%	TN%	FN%
HPNS	95	5	95	5
NB	90	10	90	10

يتضمن الجدول (7) عدة مقاييس لتصنيف وثائق الاختبار بالاعتماد على قيم جدول (4).

الجدول (7) عدة مقاييس لتصنيف وثائق الاختبار

Measure / Model	Precision	Recall	Specificity	F_1-macro	F_1-micro	
HPNS	95	95	95	95	95	95
NB	90	90	90	90	90	90

وبمقارنة نتائج المقياسين **F1-macro** و **F1-micro** للطريقة المقترحة HPNS المعتمدة على ديناميكية عمل النظام المناعي التكيفي مع نتائج الطرائق السابقة الموضحة بالجدول (8) يتبين أن أنموذجHPNS أفضل من الطرائق السابقة.

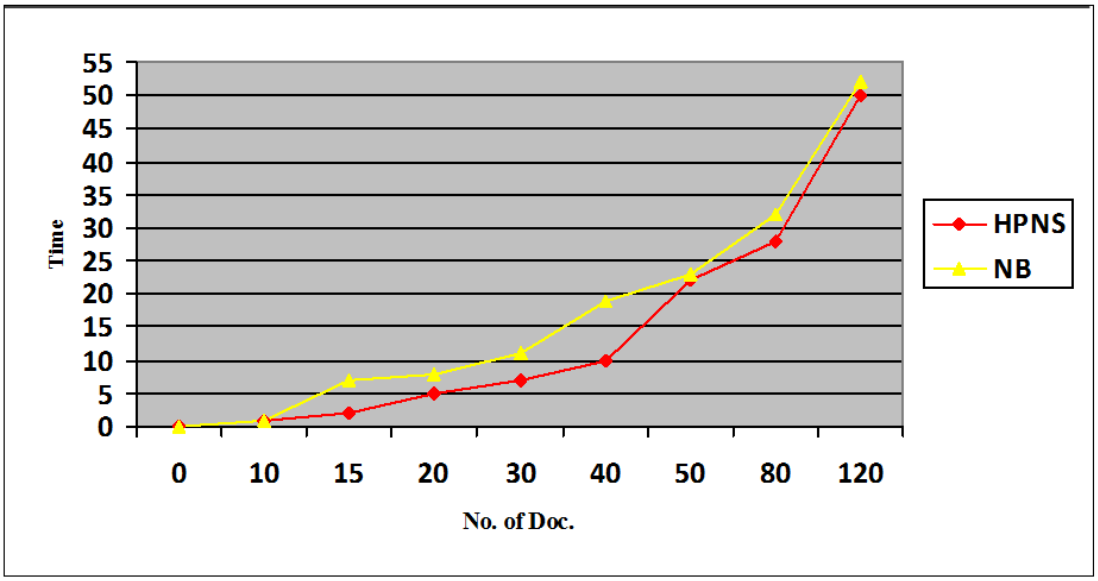
الجدول(8) مقارنة نتائج دقة مصنفات مختلفة لمقياسي F1-macro و F1-micro

Results reported by	Dataset	Classifier Used	F1-Micro	F1-Macro
[Tan et al., 2005] [36]	20 Newsgroup	Centroid	0.842 0	0.838
[Qian et al., 2007] [32]	Reuters 21578	Decision Tree	0.884	0.822
		Linear SVM	0.920	0.871
[Lan et al., 2009] [21]	Reuters 21578	SVM	0.921	0.900
		K-NN	0.840	0.825
	20 Newsgroup	SVM	0.808	0.808
		K-NN	0.691	0.691
Nada and [Rasha] 2012	Software Engineering Documents	HPNS	0.95	0.95
		NB	0.90	0.90

يعد الوقت المستغرق للتدريب والتصنيف من العوامل المهمة ومن الشكل (17) يتبين أن الوقت يتناسب طردياً مع عدد الوثائق المراد تصنيفها؛ إذ كلما زاد عدد الوثائق زاد الوقت المستغرق في التصنيف. يستغرق أنموذج HPNS لتدريب 100 وثيقة 00:01:15 ، في حين يستغرق NB 00:03:00 لتدريب الوثائق نفسها. فضلاً عن أن هيكلية وثائق هندسة البرمجيات تختلف عن وثائق الانترنيت؛ إذ تتكون وثيقة SWE من عدة صفحات لذا تستغرق وقتاً أطول من وثائق الانترنيت في التدريب والتصنيف؛ إذ عند

تدريب المصنف HPNS على 400 وثيقة انترنيت يستغرق 00:00:45 ثانية وفي التصنيف يستغرق 00:00:10 ثوانٍ في حين يستغرق NB00:01:00 دقيقة في التدريب و00:00:45 ثانية في التصنيف.

عند تصنيف وثائق متقاربة بالنوع يبرز أداء النظام المناعي العالي في التصنيف في حين يكون أداء NB أقل لكونه يعتمد على الاحتماليات و تزداد دقة NB في الأصناف المتباعدة.



الشكل (17) علاقة الوثائق مع الوقت المستغرق لتصنيفها

7. الاستنتاجات

من خلال تطبيق أنموذج HPNS في البحث لغرض تصنيف وثائق هندسة البرمجيات وعلى وفق النتائج التي تم الحصول عليها، تم التوصل إلى الاستنتاجات الآتية:بعد دراسة بعض تقانات النظام المناعي AIS اخيرت تقنية الانتقاء الايجابي والانتقاء السلبي وتهجينهما لحل المشكلة المدروسة لما لهاتين التقانتين من مميزات تساعد في عملية التصنيف. بعد ملاحظة نتائج البحث التي تم الحصول عليها تم التوصل إلى جدارة التطرق إلى الموضوع الذكائي الجديد وهو الأنظمة المناعية الاصطناعية. دقة نظام HPNS في تصنيف الوثائق عالية، حتى إن كانت أصناف الوثائق متقاربة جداً كما في وثائق هندسة البرمجيات, في حين أن طريقة NB دقتها عالية في اصناف الوثائق المتباعدة وتقل كلما كانت الأصناف متقاربة؛ إذ يحصل تداخل في أصناف الوثائق. أداء مصنف HPNS يكون نسبياً ثابتاً على مدى من أحجام مجموعة التدريب، وهذه الثبوتية تكون مفيدة؛ إذ إن نتيجة التصنيف تكون جيدة باستخدام مجموعة تدريب صغيرة الحجم وهذا ينفع المستخدمين من تدريب النظام على عدد قليل من الوثائق وفي الوقت نفسه يمتلك تنبؤات دقيقة. يمتلك النظام المناعي بصورة عامة القابلية على تكوين مصنفات ذات دقة تنبؤ أفضل باستخدام متجهات خواص ذات حجم أقل من متجهات خواص NB. إن اختيار المصنف لا يعتمد فقط على دقة التنبؤ الممكن أن يحققها، ولكن أيضاً على حجم الوقت المستغرق في التدريب؛ إذ إن المستخدم لا يرغب في انتظار النتائج أكثر من ثوانٍ قليلة وليس عدة أيام, التي تستغرقها بعض الخوارزميات. إن المصنف (HPNS) المبني على تقانة عمل النظام المناعي التكيفي يحقق دقة تنبؤ جيدة وفي وقت قليل جداً يسمح باستخدامه في مشاكل العالم الحقيقي.

المصادر

[1] Aickelin, U., (2004), "Artificial Immune Systems (AIS) – A New Paradigm for Heuristic Decision Making", Inciter Keynote talk, Annal Operational Research conference 46, York, UK.

[2] Brownlee, J., (2008), "Clonal Selection as an Inspiration for Adaptive and Distributed Information Processing", PhD thesis, Swinburne University, Melbourne, Australia.

[3] Burnet, F., M., (1959), "The Clonal Selection Theory of Acquired Immunity ". Vanderbilt, University, Press, Nashville, T.N. Canada, Journal of Information Technology Education Volume 6.

[4] Cohen, W., Singer, Y., (1999), "Context sensitive learning methods for text categorization", ACM Transactions on Information Systems, vol. 17, no. 2, pp. 141- 73.

[5] De Castro L.N., Von Zuben, F.J., (2002), "Learning and optimization using the clonal selection principle", IEEE Trans. Evol. Comput., vol. 6, no. 3, pp 239—251.

[6] De Castro, L. N., Timmis, J., (2002). "Artificial Immune Systems: A New Computational Intelligence Approach", ISBN 1-85233-594-7, Springer, England.

[7] Engelbrecht, Andries, P., (2007), "Computational Intelligence: An Introduction", 2nd edition, University of Pretoria, South Africa, John Wiley & Sons Ltd.

[8] Forman, George, Eshgi, Kave, Chiocchetti, (2005), "Finding Similar Files in Large Document Repositories", Proc. 11th ACM International Conf. on Knowledge Discovery and Data Mining (KDD'05), 21-25, Chicago, Illinois, USA

[9] Forrest, S., Perelson, A., Allen, L., Cherukuri, R., (1994), "Self-Nonself Discrimination in a Computer", Proc. of the IEEE Symposium.

[10] Fouad, Walid, A., Badr, Amr, A., Abdel- Rahman, Ebrahim, F., (2007), "A compartive Study of web document classification approaches", Proc. 37th International Conf. on computers and industrial Engineering, pp. 197-206, Alexandria, Egypt.

[11] Goldsby, Richard, A., Kindt, Thomas, J., Osborne, Barbara, A, Kuby, J., (2003), " Immunology", 5th Edition, W. H. Freeman and Company.

[12] Greensmith, J., Aickelin, U., (2009), "Artificial Dendritic Cells: Multi-faceted Perspectives", in Human-centric information processing through granular modeling, studies in computational intelligence (182). Springer, Berlin, pp 373395, ISBN 9783540929154.

[13] Greensmith, J.,(2003), "New Frontiers For An Artificial Immune System", MSC Thesis, University of Leeds, Hewlett Packard Labs Technical Report Number HPL.

[14] Hotho, A., Nurnberger, A., PaaB, G., (2005), " A Brief survey of Text Mining", Journal for Computational Linguistics and Language Technology, Vol 20, pp. 19-62.

[15] Ikonomakis, M., Kotsiantis, S. , Tampakas, V., (2005), "Text Classification Using Machine Learning Techniques", WSEAS Transactions on Computers, Issue 8, Volume 4, pp.966-974.

[16] Iqbal, A., (2006), "Danger Theory Metaphor In Artificial Immune System For System Call Data", PhD theises, Universiti Teknologi Malaysia.

[17] Ismail, Nabil, A., Abdul-Kader, H., Al-Sheshtawi, Khaled, A.,(2010), "Artificial Immune Clonal Selection Classification Algorithms for Classifying Malware and

Benign Processes Using API Call Sequences", IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.4, pp.31-39.

[18] Joachims, T., (1998), " Text categorization with support vector machines: Learning with many relevant features". Proc. Of European Conf. on Machine Learning (ECML), Vol 1398, pp. 137-142.

[19] Kamthan,P., (2007), "On the Prospects and Concerns of Integrating Open Source Software Environment in Software Engineering Education", Concordia University, Montreal, Quebec,

[20] Khelil, H., Benyettou, A.,(2006), Artificial Immune Systems For Illnesses D Diagnostic, Ubiquitous Computing and Communication Journal, Algeria.

[21] Kuby, J., (1997), "Immunology", 3rd Ed., W. H. Freeman and Co.

[22] Lan, M., Tan, C. L., Su. J., and Lu, Y.2009. Supervised and Traditional Term Weighting Methods for Automatic Text Categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 31 (4), pp. 721 – 735.

[23] Matzinger, P., (1994), "Tolerance, danger and the extended family", Annual Reviews in Immunology,12:991–1045.

[24] Meyer, B.J., Meij, H.S., Grey, S.V., Meyer, A.C., (1996), "Fisiologie van die mens –Biochemiese", fisiese en fisiologiese begrippe. Kagiso Tersier - Cape Town, 1st Edition.

[25] Middlemiss, M., (2006), "Positive and negative selection in a Multilayer Artificial Immune System", information science paper No. 2006/03 Dunedin University of Otago, NEW ZEALANDPP. 17.

[26] Mitchell, T., (1997), "Machine Learning". McGraw Hill, New York.

[27] Mohammad, A., (2008), "Support Vector Machine Text Classification for Arabic Articles: Ant Colony Optimization Based Feature Subset Selection", PhD thesis, Arabic Academy for Banking and Financial Sciences, Amman, Jordan.

[28] Nanda, Satyasai, J., (2009), "Artificial Immune Systems: Principle, Algorithms And Applications", MSC thesis, Thapar University, India.

[29] Negi, P., (2006), "Artificial Immune System Based Urban Traffic Control", MSC thesis, A&M University, Texas.

[30] Ojasvini, Nitesh, Piyush, Thakur, N., Rehalia, A., (2018), "Intrusion Detection System Using Artificial Immune System: A case study ", International Journals Advanced Research in computer Science and Software Engineering, ISSN: 2277-128x, vol. 8, Isuue -2.

[31] Onan, A., (2015), " Artificial Immune System Based Web Page Classification", In: Silhavy R., Senkerik R., Oplatkova Z., Prokopova Z., Silhavy P. (eds) Software Engineering in Intelligent Systems. Advances in Intelligent Systems and Computing, vol 349. Springer, Cham".

[32] Özgür, A., (2004), "Supervised and Unsupervised Machine Learning Techniques For Text Document Categorization", MSC thesis, , Boˇgazici University.

[33] Pawar, P., Gawande, S., (2012), " A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4.

[34] Pazzani, M., Billsus, D., (1997), "Learning and revising user profiles: the identification of interesting web site ", machine learning, pp. 313-331.

[35] Qian, T., Xiong, H., Wang, Y., and Chen, E. 2007. On the strength of hyperclique patterns for text categorization. An International Journal Information Sciences, Vol. 177, pp. 4040–4058.

[36] Ramdane, ch., Chikhi, S., (2017), "Negative Selection Algorithm: Recent Improvements and Its Application in Intrusion Detection System", International Journal of Computing Academic Research (IJCAR), ISSN 2305-9184, vol.6, no. 2,pp. 20-30

[37] Romero, A., Niño, F., (2007), "An Artificial Immune System Based on Information Theory for Keyword Extraction from Text Documents", Revista Avances en Sistemas e Informática, Vol.4 No. 2, pp. 25-32, Edición Especial: II Congreso Colombiano de Computación - CCC.

[38] Saranya, C., Thenmozhi, D., (2015), " Machine Learning Approach to Document Classification using Concept based Features ", International Journal of Computer Applications (0975 – 8887) Volume 118 – No.20.Sciences, Amman, Jordan.

[39] Sebastiani, F., (2002), "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, no. 1, pp. 1-47.

[40] Semberecki, P., Maciejewski, H., (2017), " Deep Learning methods for Subject Text Classification of Articles", Proceedings of the Federated Conference on Computer Science and Information Systems pp. 357–360 ISSN 2300-5963 ACSIS, Vol. 11.

[41] Songbo, T., Cheng, X., Ghanem, M. M., Wnag, B., and Xu, H. 2005. A novel refinement approach for text categorization. In the Proceedings of Fourteenth ACM International Conference on Information and Knowledge Management, pp 469 – 476.

[42] Tan, S., (2008), "An improved centroid classifier for text categorization". Journal of Expert System with Applications, Vol 35, pp 279 – 285.

[43] Twycross, Jamie, P., (2007)," Integrated Innate And Adaptive Artificial Immune System Applied To Process Anomaly Detection", PhD theises, University of Nottingham, UK.

[44] Wei , T., (2010), "Homology Modeling of Toll-Like Receptor Ligand-Binding Domains: A Leucine-Rich Repeat Assembly Approach",PhD theises, Ludwig-Maximilians-Universität München.

[45] Williams, C., Harry, R., and McLeod, J. (2007), "Mechanisms of apoptosis induced DC suppression", Journal of Immunology, 120S158.