Syrian SMS Spam Classification

نهدف في هذا المشروع الى بناء نظام ذكي يتمثل عملهُ الأساسي بتصنيف الرسائل النصية SMS الى نوعين من الرسائل :

1. رسائل إعلانية (SPAM) .
2. رسائل طبيعية بين المستخدمين (HAM) .

يكمن السبب الذي دعا الى بناء هذا النظام في انهُ بعض المستخدمين للهواتف النقالة يعانون من الرسائل الاعلانية التي تُرسل إليهم وتكون بمثابة مصدر ازعاج لهم .

مرحلة البناء الأساسية لهذا النظام تتمثل بالعمل على تصنيف الرسائل المدخلة اليهِ الى الصنفين المذكورين سابقاً فقط ,

بحيث من الممكن في مراحل متقدمة من العمل تطوير هذا النظام لكي يصبح تطبيق متكامل يمكن تشغيلهُ على الهواتف النقالة .

النظام يتعامل مع الرسائل النصية TEXT MESSAGES (SMS) العربية (السورية) .

REFERENCE STUDY ON WAYS TO SOLVE THE ISSUE

يوجد في العديد من الهواتف الذكية تقنيات تمكن المستخدمين من اضافة رقم معين إلى لائحة Blacklist,

بحيث يتم حجب جميع الرسائل القادمة من الأرقام ضمن هذه اللائحة.

لكن مع تجريب هذه التقنية مع الرسائل الاعلانية , لاحظنا عدم تمكنها من حجب الرسائل عن المستخدم بسبب عدم وجود رقم واحد, أو في أغلب الاحيان لا يوجود رقم من اساساً.

بالإضافة إلى ضرورة إضافة الأرقام يدوياً الى لائحة الحظر .

ولحل هذه المسألة تم الاعتماد على استخدام الكثير من الطرق, نذكر منها :

- Artificial Immune System
- Support Vector Machine
- Winnow algorithm
- Content based using Bayesian filtering techniques
- feature based and compression-model based filters
- Content based and challenge-response
- Behavior based social network and temporal analysis

بعد الاطلاع على عدة أوراق بحثية حول الموضوع المدروس , تم الاعتماد على مجموعة من الخطوات والخوارزميات في كل مرحلة من مراحل تطوير النظام .

بحيث معظم الأبحاث التي تدور حول الموضوع المدروس كانت تستخدم خطوات الأساسية التالية :

1. (Data Acquisition) .
2. (Feature Extraction) .
3. (Feature Selection) .
4. (Classification Module) .

بحيث وجدنا من التقنيات المستخدمة ضمن مرحلة ال Feature Extraction : (Bag Of Word (BOW) , Tf-Idf)

والتقنيات المستخدمة ضمن مرحلة ال Classification Module : svm , naïve bayes والشبكات العصبونية .

في الحل الخاص بنا , قمنا بإضافة خطوة مهمة جداً قبل استخلاص الميزات (features) من البيانات , بحيث هذه الخطوة هي text preprocessing , والتي تتمثل بإجراء عدة عمليات معالجة نصية على الرسائل قبل مرحلة استخلاص الميزات من ضمن نصوص هذه الرسائل .

وبذلك تم تقسيم مراحل تطوير النظام الى 5 مراحل أساسية :

1. الحصول على البيانات اللازمة لبناء النظام وتدريبهُ (Data Acquisition) .
2. تجهيز النصوص الخاصة بكل الرسائل ضمن بيانات التدريب (Text Preprocessing) .
3. استخراج الميزات من ضمن نصوص الرسائل الخاصة ببيانات التدريب (Feature Extraction) .
4. اختيار افضل الميزات المستخرجة من نصوص الرسائل (Feature Selection) .
5. بناء نموذج التصنيف وتدريبهُ بعد تجهيز بيانات التدريب لهُ (Classification Module) .

بحيث تم اختبار عدة تقنيات في كل مرحلة من المراحل السابقة .

EXPLAINING THE SOLUTION STAGES

الان سوف نقوم بشرح كل مرحلة من المراحل الأساسية السابقة بالتفصيل .

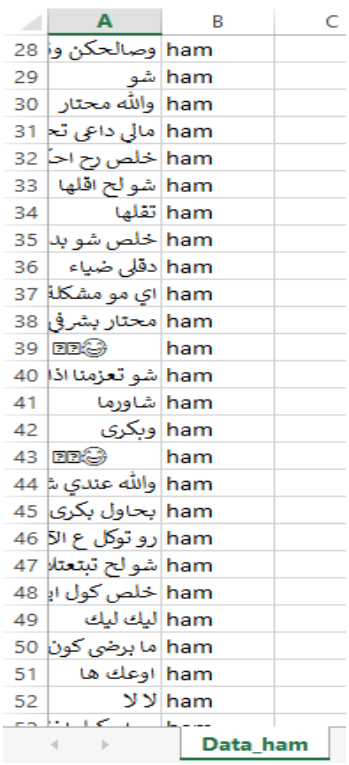
Data Acquisition

تم تجميع الرسائل من الهواتف النقالة الخاصة بأعضاء الفريق بالإضافة الى الأقارب والاصدقاء, وذلك من خلال طريقتين :

1) **الطريقة الأولى :** تم استخلاص الرسائل الاعلانية والرسائل الطبيعية وترتيبها ضمن ملفات CSVمع ارفاق الصنف الخاص بكل رسالة نصية من الرسائل , أي الملف يحتوي على عمودين :
   1. العمود الأول : يحتوي على نص الرسالة .
   2. العمود الثاني : يحتوي على صنف الرسالة (Spam , Ham) .

بحيث قمنا بتصنيف الرسائل يدوياً , ووضعها ضمن الملفات .

الصورة التالية توضح شكل الملف الحاوي على الرسائل :



	A	B	C
28	وصالحكن و	ham	
29	شو	ham	
30	والله محتار	ham	
31	مالي داعي تح	ham	
32	خلص رح احـ	ham	
33	شو لح اقلها	ham	
34	تقلها	ham	
35	خلص شو بد	ham	
36	دقلي ضياء	ham	
37	اي مو مشكلة	ham	
38	محتار بشرفي	ham	
39	😂😂	ham	
40	شو تعزمنا اذا	ham	
41	شاورما	ham	
42	وبكرى	ham	
43	😂😂	ham	
44	والله عندي ش	ham	
45	بحاول بكرى	ham	
46	رو توكل ع الـ	ham	
47	شو لح تبتعتلا	ham	
48	خلص كول ايـ	ham	
49	ليك ليك	ham	
50	ما برضى كون	ham	
51	اوعك ها	ham	
52	لا لا	ham	

Data_ham

	A	B	C
1	تم الغاء باقة	spam	
2	تم الغاء باقة	spam	
3	عذرا لا يمكن	spam	
4	عذرا لم يتم ت	spam	
5	دقيقة خليوبا	spam	
6	عذرا لم يتم ت	spam	
7	عذرا باقة ياه	spam	
8	تم الغاء باقة	spam	
9	عذرا لم يتم ت	spam	
10	عذرا هذه البا	spam	
11	عذرا باقة ياه	spam	
12	عذرا لا يمكن	spam	
13	باقات جديدة	spam	
14	لتستفيد طول	spam	
15	باقات يوم الـ	spam	
16	اطلب هلأ 9	spam	
17	لعشاق كرة الـ	spam	
18	كل شي بيتعلم	spam	
19	شركة الهرم تر	spam	
20	خدمة أدب با	spam	
21	حمل رنة بغن	spam	
22	لقد تم ضبط	spam	
23	حمل رنة بغن	spam	
24	سيتم إلغاء خ	spam	
25	لقد تم ضبط	spam	

Data_spam

2) **الطريقة الثانية** : تم استخراج الرسائل من خلال استخدام برنامج (SMS Backup and Restore) , والذي يعطي الرسائل مع الخصائص المتاحة ويخزنها ضمن ملف XML , وتم الاعتماد على هذه الملفات من اجل استخراج الرسائل منها ,

اعتمدنا على خصائص جهة الاتصال من اجل التعرف على صنف الرسالة وتخزينهُ دون الحاجة الى تصنيفها يدوياً كما في الطريقة السابقة , بحيث إذا كانت جهة الاتصال هي (Unknown) فهذا يعني انها رسالة إعلانية ويكون الصنف المقابل لها هو (spam) , والا سوف تكون رسالة عادية لان الرسائل العادية تحمل إما اسم او رقم المرسل .

بعد تجميع الرسائل مع اصنافها ضمن ملفات باستخدام الطريقتين السابقتين , نقوم الان بوضع هذه الرسائل ضمن **Pandas Data frame ,** وقمنا بإجراء العمليات التالية :

1. التخلص وحذف الرسائل المتكررة .
2. ترتيب الرسائل بحسب طول النص الخاص بكل رسالة .
3. المساواة بين عدد الرسائل الاعلانية (spam) وعدد الرسائل العادية (ham) وذلك من خلال التخلص من الرسائل القصيرة جداً .

قمنا بإجراء هذه الخطوات بهدف تحسين أداء المصنف الذي نسعى لتدريبهُ على هذه البيانات .

Text Preprocessing

في هذه المرحلة , قمنا بتطبيق عدة عمليات معالجة بهدف تحسين النصوص الخاصة بالرسائل وتوحيدها , بحيث قمنا بتطبيق العمليات التالية على النص :

- إزالة الروابط في حال وجدت ضمن نص الرسالة .
- **Text Normalization** : توحيد شكل النص وذلك من خلال توحيد الاشكال المختلفة لبعض الاحرف الى شكل واحد فقط .
- ازالة الاحرف المتكررة .
- إزالة الاحرف الأجنبية .
- تجزيع الكلمات .
- إزالة الأرقام .
- إزالة الكلمات التي لا تضيف معنى مفيد (Stop Words) .
- إزالة محارف معينة مثل (.....[]\|/) وهكذا .

feature Extraction

في هذه المرحلة قمنا بالاعتماد على تقنيتي (**Bag Of Word (BOW) , Tf-IDf**) والتي قمنا بتعلمهم ضمن جلسات العملي لهذه المادة ,

وذلك بهدف استخلاص الميزات المهمة من نصوص الرسائل النصية والتي سوف تساعد على تصنيف الرسائل بشكل صحيح بالاعتماد على هذه الميزات .

قمنا بتعيين المجال المأخوذ لعدد الكلمات المترابطة والمتتالية على انهٔ n = 1,2,3 , أي :

- **unigram : n = 1**
- **bigram : n = 2**
- **trigram : n = 3**

Feature Selection

بعد استخلاص الميزات من ضمن نصوص الرسائل , سوف نقوم باختيار الميزات (features) الأفضل من بينها والتي سوف تزيد من أداء هذا النظام وتعطي نتائج افضل .

بحيث قمنا بالاعتماد على تقنية **Chi2** :

- والتي تعتبر تقنية إحصائية تقيس الارتباط بين اثنين من المتغيرات التابعة لصنفين مختلفين.

- و تستخدم هذه التقنية لتحديد ما إذا كان هناك فرق ذي دلالة إحصائية (أي اختلاف واضح ليس فقط بسبب الصدفة) بين الترددات المتوقعة والترددات المرصودة في فئة أو أكثر من الحالات المدروسة.

- أي تمكننا من تحديد الميزات (features) ذات أعلى قيمة إحصائية من بين بيانات التدريب.

- بحيث يتم القياس بالاعتماد على الفرق بين المتغيرات العشوائية .

- بمعنى اذا تم استخدام هذه التقنية سوف يتم إزالة الميزات (features) التي من المرجح ان تكون مستقلة عن الصنف وبالتالي لا صلة لها بالتصنيف .

بعد تطبيق هذه التقنية على الميزات المستخرجة من قِبل تقنيتي (**Bag Of Word (BOW) , Tf-Idf**) يتم حساب القيمة Score لكل ميزة ونأخذ اعلى 10% من هذه الميزات .

Classification Module

بعد استخلاص الميزات (features) , نقوم الان بتدريب المصنف الخاص بنا بهدف يصبح قادر على تصنيف الرسائل الى رسائل إعلانية ورسائل عادية , وذلك بالاعتماد على الميزات المستخلصة ضمن المراحل السابقة .

قمنا بالاعتماد على ثلاثة أنواع مختلفة من نماذج التصنيف , وهي :

1. **Naive Bayes Classifier** :
   - عبارة عن مصنف احتمالي بسيط يستند الى تطبيق نظرية بايز مع افتراضات استقلالية قوية بين الميزات (features) المستخلصة .
   - بحيث يعتبر من الطرق الشائعة والمستخدمة بكثرة في تصنيف النصوص .
   - وهذا السبب الذي دفعنا الى استخدامهُ وتدريبهُ على البيانات (الميزات المستخلصة والمختارة) الخاصة بنا .
2. **SVM (Support Vector Machine)** :
   - هو نموذج تعلم تحت اشراف (Supervised Learning) مع خوارزميات التعلم والتي تعمل على تحليل البيانات المستخدمة في عمليات التصنيف Classification .
   - عند إعطاء مجموعة من الأمثلة التدريبية ، يتم تمييز كل منها على أنه ينتمي إلى فئة واحدة أو أخرى من فئتين .
   - تقوم خوارزمية تدريب SVM بإنشاء نموذج يعين أمثلة جديدة لفئة أو لأخرى ، مما يجعله مصنفًا خطيًا ثنائيًا غير محتمل .
   - نموذج SVM هو تمثيل للأمثلة كنقاط في الفضاء ،معيّن , بحيث يتم تقسيم أمثلة الفئات المنفصلة على فجوة واضحة واسعة بقدر الإمكان.
   - ثم يتم تعيين أمثلة جديدة في نفس المساحة ويتوقع أن تنتمي إلى فئة تستند إلى جانب الفجوة التي تقع عليها.
   - وبذلك يعتبر هذا النموذج مصنفاً جيداً مما دفعنا الى تجربتهُ على البيانات الخاصة بنا , بحيث قمنا باختيار نواة Kernel من نوع Linear .

3. ANN (Artificial Neural Network) :

- قمنا ايضاً بالاعتماد على تدريب شبكة عصبونية مؤلفة من طبقتين خفيتين .
- عدد العقد في كل طبقة هو 50 عقدة .
- مع نسبة Dropout تبلغ 50%.
- قمنا باستخدام توابع التنشيط Activation function التالية :
  1) **relu** : في الطبقات المخفية .
  2) **Sigmoid** : في طبقة التصنيف .

RESULTS

الان سوف نقوم بعرض نتائج اختبار التقنيات السابقة بتراكيبها المختلفة ,

بحيث سوف نعتد على المقاييس التالي من اجل تقييم الأداء :

1. **Precision**
2. **Recall**
3. **Accuracy**

نعتبر هنا الرسائل الاعلانية (spam) هي الPositive , وتكون لدينا الحالات التالية :

- **True Positives (TP)** : عندما تكون الرسائل هي رسائل اعلانية (spam) ونقوم بتصنيفها على أنها رسائل اعلانية (spam).
- **True Negatives (TN)** : عندما تكون الرسائل هي رسائل عادية (ham) ونقوم بتصنيفها على أنها رسائل عادية (ham) .
- **False Positives (FP)** : عندما تكون الرسائل هي رسائل عادية (ham) ونقوم بتصنيفها على انها رسائل اعلانية (spam).
- **False Negatives (FN)** : عندما تكون الرسائل هي رسائل اعلانية (spam) ونقوم بتصنيفها على أنها رسائل عادية (ham) .

يتم حساب الPrecision , والRecall كما يلي :

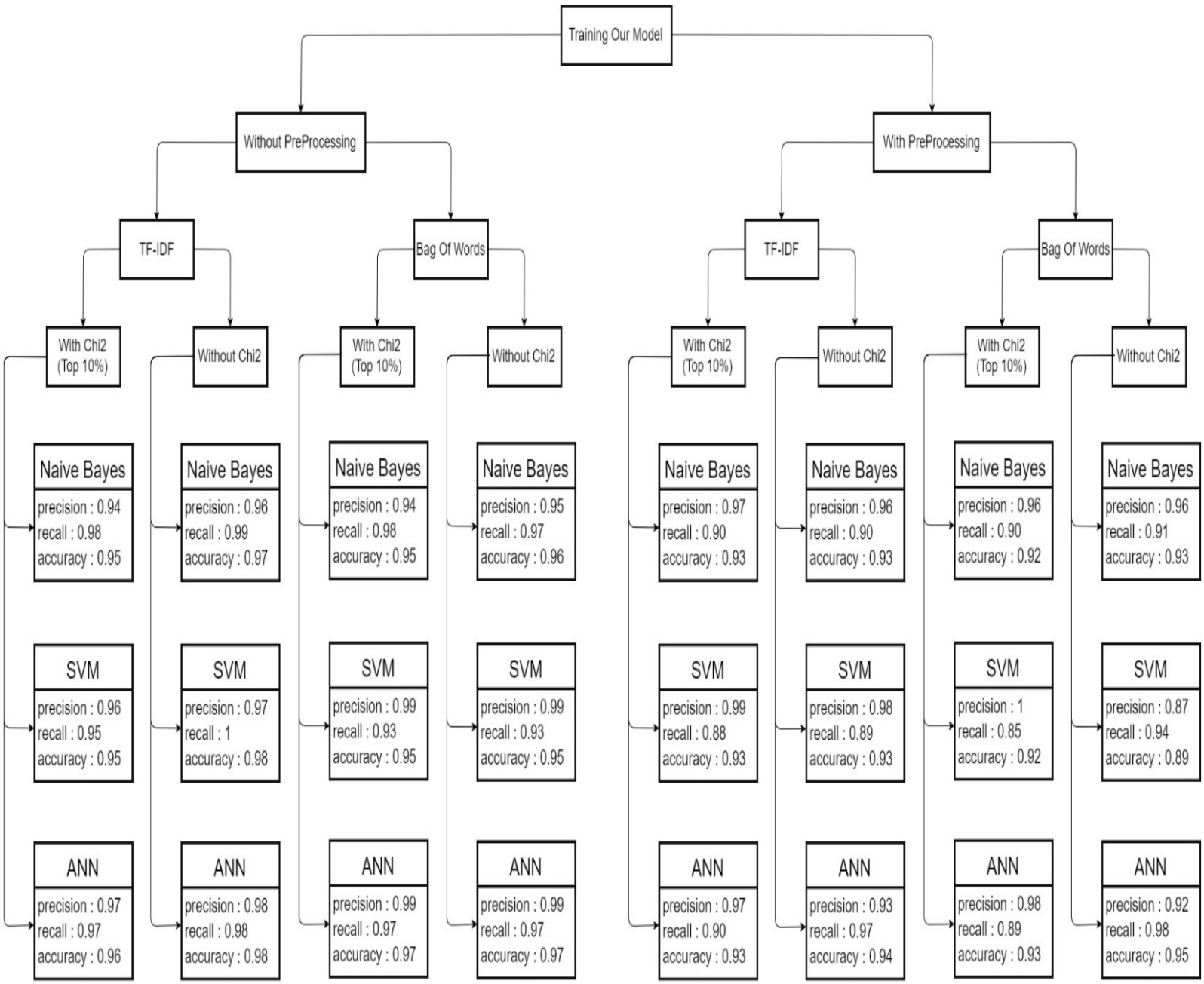
$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

واخيراً قمنا بالاعتماد على النموذج الفكري التالي:

في حال تم تصنيف بعض رسائل (spam) على أنها (ham) (FN) , لن يشكل ذلك مشكلة كبيرة , لكن في حال تم تصنيف بعض رسائل (ham) على أنها (spam) , من الممكن ان يشكل ذلك مشكلة (FP), لذلك نحاول في مسألتنا رفع نتيجة ال precision قدر الإمكان, وكانت طريقة من طرق رفعها هي ال Preprocessing و Feature Selection .

توضح الشجرة التالية , آلية تدريب المصنف الخاص بنا باستخدام التقنيات المذكورة سابقاً مع افضل النتائج التي تم الحصول عليها :



REFERENCES

1) An Analysis of Various Algorithms for Text Spam Classification and Clustering Using RapidMiner and Weka .

By : Kamahazira Zainal & Zalisham Jali

Link :

https://www.researchgate.net/publication/277564480_An_Analysis_of_Various_Algorithms_For_Text_Spam_Classification_and_Clustering_Using_RapidMiner_and_Weka

2) Filtering Spam E-Mail from Mixed Arabic and English Messages : A Comparison of Machine Learning Techniques .

By : Alaa Mustafa El-Halees

Link :

https://www.researchgate.net/publication/220413606_Filtering_Spam_E-Mail_from_Mixed_Arabic_and_English_Messages_A_Comparison_of_Machine_Learning_Techniques

3) The Impact of Feature Extraction and Selection on SMS Spam filtering

By : Alper Kursat Uysal

Link :

https://www.researchgate.net/publication/236868701_The_Impact_of_Feature_Extraction_and_Selection_on_SMS_Spam_Filtering

4) Chi – Square Test for feature Selection in Machine Learning

By : Sampath Kumar Gajawada

Link :

https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

5) NLTK (Natural Language ToolKit) 3.4.5 documentation

Link :

https://www.nltk.org/

6) Scikit – Learn

Link :

https://scikit-learn.org/stable/

THE END

-.-