Ministry of Higher Education
& Scientific Research
*University of Mosul*
*College of Computer Sciences*
*and Mathematics*

# Data Mining for Classification Mobile Phone Messages

**Hind Salman Hassan**

**M.Sc./Thesis**
**Computer Science**

**Supervised By**
**Dr.Ghayda Abdul-Aziz Majeed**

**Assist Prof.**

**2014 A.D** **1435 A.H**

# Abstract

SMS classification technology has an important significance to assist people in dealing with SMS messages. Although SMS classification can be performed with little or no effort by people, it still remains difficult for computers. Machine learning offers a promising approach to the design of algorithms for training computer programs to efficiently and accurately classify short text message data.

A corpus of 4000 SMS messages were manipulated by performing some preprocessing methods on them; those methods are lexical text analysis, replacement of abbreviations and stop words removal. About (1620) terms of abbreviation are collected and (360) terms of stop words were used.

Then, features extraction was applied on the messages in order to collect the important features for each message. Those features used to reduce the dimensionality of the data under processing. There are three approaches used for this reasons, which are parts of speech tagging, stemming and term frequency.

In order to assign weight to each term in the SMS, term frequency-inverse document frequency (TF-IDF) technique was used, which has an

effective role in the enhancement of the performance of the proposed classification system.

Two swarm optimization algorithms have been used in the classification of SMS messages_ cat swarm optimization (CSO) and particle swarm optimization (PSO) algorithms.

Finally, the performance of the classification methods used was evaluated by using a number of measures, which are precision, recall, F-score accuracy, and error rate measures.

Experimental results show that classifiers that have been used in this thesis are understandable, accurate and perform well. The system was implemented using Microsoft visual c# 2008.

وزارة التعليم العالي والبحث العلمي
جامعة الموصل
كلية علوم الحاسوب والرياضيات


# التنقيب في البيانات لأجل تصنيف رسائل الهاتف النقال

## هند سلمان حسن

رسالة ماجستير
علوم الحاسوب

بإشراف
د.غيداء عبد العزيز مجيد
أستاذ مساعد

١٤٣٥هـ ٢٠١٤م

# الملخص


ان لتقنية تصنيف رسائل الهاتف المحمول النصيةاهمية كبيرة في مساعدة الناس على التعامل مع تلك الرسائل. فعلى الرغم من ان تصنيف الرسائل النصية القصيرة يمكن القيام به مع قليل من الجهد من قبل الانسان، الا ان ذلك صعبا على أجهزة الحاسوب القيام به. حيث قدم موضوع تعليم الآلة نهجا واعدا في تصميم خوارزميات لتدريب الحاسوب على تصنيف رسائل الهاتف بدقة وكفاءة.

تم في هذا البحث استخدام مجموعة من رسائل الهاتف النصية القصيرة, تكونت من (4000) رسالة كذخيرة للعمل حيث تم تهيأتها للمعالجة بواسطة أساليب تحليل النص المعجمية، واستبدال المختصرات وازالة كلمات التوقف. حيث تم جمع حوالي (١٦٢٠) مصطلح من المختصرات الشائعة وحوالي (٣٦٠) كلمة توقف.

تضمنت المرحلة التالية استخلاص الميزات من الرسائل النصية من أجل جمع السمات الهامة لكل رسالة، هذه المرحلة استخدمت للحد من أبعاد البيانات قيد المعالجة، وقد تم لأجل ذلك استخدام ثلاثة تقنيات وهي وسم أجزاء الكلام، وعملية ارجاع الكلمات الى جذورها(stemming) و حساب تردد المصطلحات.

تم استخدام تقنية TF-IDF لغرض تحديد الاوزان لكافة المصطلحات في SMS، والتي أثبتت فعالية كبيرة في تعزيز أداء نظام التصنيف المقترح.

استخدمت خوارزميتان من خوارزميات السرب الامثلية في تصنيف الرسائل القصيرة، وهي خوارزمية سرب القطط (CSO) وخوارزمية سرب الطيور(PSO) , ومن ثم تم تقييم أداء طرق التصنيف المستخدمة بواسطة عدد من المقاييس, و هي الدقة والاسترجاع و دقة F والكفاءة ونسبة الخطأ.

أظهرت النتائج التجريبية أن المصنفات التي استخدمت في هذه الرسالة واضحة و دقيقة و كان اداؤها جيدا.استخدمت لغة #c في برمجة النظام المقترح.