ARABIC TEXT CLASSIFICATION BASED ON ARTIFICIAL BEE COLONY ALGORITHM AND SEMANTIC RELATIONS

BY

MUSAB MUSTAFA HIJAZI

A thesis submitted in fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science

Kulliyyah of Information and Communication
International Islamic University Malaysia

AUGUST 2022

ABSTRACT

Documents contain a tremendous quantity of important human information. The use of automatic text classification is necessitated by the substantial increase in the volume of machine-readable documents for public or private access. Text classification is the process of categorizing or organizing documents into a predetermined set of classes. Western languages, namely English, have received a lot of attention, whereas the Arabic language has received far less attention. Arabic text categorization methods emerged spontaneously as a result of the vast volume of diverse textual material provided in Arabic on the internet. The selection of features is an essential step in text categorization. It is an important preprocessing approach for effective data analysis, in which just a subset of the original data features is chosen after eliminating noisy, unnecessary, or duplicated features. Bag of Words (BoWs) representation is considered the simplest representation of texts. Most Arabic researchers have been trying to find an accurate Arabic text classification based on the traditional Bag of Words (BoWs) for data representation which does not consider the semantic relationships between the words, such as synonymy and hypernyms. This research aims to build a model for Arabic text classification using the Artificial bee colony algorithm as a feature selection method and Arabic WordNet (AWN) as a lexical and semantic resource to utilize the semantic relationships between the words. The results of the research showed that the proposed Chi-square – Binary Artificial Bee Colony chi-BABC feature selection method was able to reduce the dimensionality of the feature set and at the same time improve the text classification. It was able to reduce approximately 89% of the original feature list size when the Naïve Bayes classifier was used as a fitness function. On the other hand, around 94% of the original feature list size was reduced by the proposed feature selection method when Support Vector Machines was utilized as a fitness function. The proposed FS method was evaluated using Support Vector Machine, C4.5 Decision tree, and Naïve Bayes. Experiments showed that the proposed FS improved the performance of Arabic Text Classification with superior results for SVM with 86.9% compared with 84.5, and 77.3 for NB, and C4.5 respectively. Furthermore, the proposed FS method was compared with PSO, ACO, and GA. The experiment results showed that the proposed method outperformed the others by having 86.9% compared with 84.7%, 83.4%, and 82.7 for PSO, ACO, and GA respectively. Finally, utilizing concepts and semantic relations between them enriches the text representation by adding more semantic meaning, improving the text classification performance. The text classification performance based on grouping methods was enhanced by 2% for category term relation and 2%, and 3% for related to and has holo member relations respectively. The best classification performance was when the holo member relation is part of combined relations. The superior text classification result was 81.2 for the combination of related-to with has holo member relations while the lowest result was 78.6 for the combination of has hyponym with category term relations.

ملخص البحث

تحتوي المستندات على كمية هائلة من المعلومات المهمة.لذا فإن استخدام التصنيف التلقائي للنص أمر ضروري بسبب الزيادة الكبيرة في حجم المستندات المقروءة آليًا للاستخدام العام أو الخاص. تصنيف النص هو عملية تبويب النصوص أو تنظيمها في مجموعة فئات محددة مسبقًا. لقد حظيت اللغات الغربية وخاصة اللغة الإنجليزية باهتمام كبير بينما لم تحظ اللغة العربية باهتمام مماثل. ظهرت طرق لتصنيف النص العربي نتيجة للكم الهائل من المواد النصية المتنوعة المتوفرة باللغة العربية على الإنترنت. ويعد اختيار الكلمات خطوة أساسية في تصنيف النص. حيث تعتبر طريقة من طرق المعالجة المسبقة للنصوص وهي مهمة لتحليل البيانات بشكل فعال، حيث يتم اختيار مجموعة جزئية فقط من مجموعة الكلمات الأصلية بعد التخلص من الكلمات غير الضرورية أو المكررة. يعتبر نموذج حقيبة الكلمات (Bag of Words (BoWs أبسط تمثيل للنصوص. يحاول معظم الباحثين العرب إيجاد تصنيف دقيق للنص العربي بناءً على التمثيل التقليدي للكلمات (BoWs) والذي لا يأخذ في الاعتبار العلاقات الدلالية بين الكلمات، مثل المترادفات والأسماء الشاملة. يهدف هذا البحث إلى بناء نموذج لتصنيف النص العربي باستخدام خوارزمية مستعمرات النحل الصناعية كأسلوب لاختيار الكلمات التي ستستخدم في تمثيل النص و WordNet العربية ((AWN كمصدر معجمي ودلالي للاستفادة من العلاقات الدلالية بين الكلمات. أظهرت نتائج البحث أن طريقة الاختيار المقترحة باستخدام اختبار مربع كاي مع مستعمرة النحل الاصطناعية الثنائية chi-BABC كانت قادرة على تقليل عدد الكلمات المستخدمة في تمثيل النص وفي نفس الوقت تحسين تصنيف النص. حيث كانت طريقة اختيار الكلمات المقترحة قادرة على تقليل ما يقرب من 89٪ من حجم قائمة الكلمات الأصلية عندما تم استخدام مصنف Naïve Bayes كدالة كفاءة. من ناحية أخرى، تم تقليل حوالي 94٪ من حجم قائمة

الكلمات الأصلية من خلال طريقة اختيار الميزة المقترحة عندما تم استخدام شعاع الدعم الآلي SVM كدالة كفاءة. تم تقييم طريقة اختيار الكلمات المقترحة باستخدام شعاع الدعم الآلي وشجرة القرار C4.5و Naïve Bayes. أظهرت التجارب أن طريقة اختيار الكلمات المقترحة حسّنت أداء تصنيف النص العربي مع نتائج متفوقة لـ SVM بنسبة 86.9٪ مقارنة بـ 84.5 و 77.3 لـ NB و C4.5 على التوالي. يالإضافة إلى ذلك، تمت مقارنة طريقة اختيار الكلمات المقترحة اعتمادا على خوارزمية النحل مع PSO و ACO و GA. أظهرت نتائج التجربة أن الطريقة المقترحة تفوقت على الطرق الأخرى بنسبة 86.9٪ مقارنة بـ 84.7٪ و 83.4٪ و 82.7 لكل من PSO و ACO و GA على التوالي. أخيرًا، أدى استخدام المفاهيم والعلاقات الدلالية بينها إلى إثراء تمثيل النص عن طريق إضافة المزيد من المعنى الدلالي، وتحسين أداء تصنيف النص. كما تم تحسين أداء تصنيف النص المعتمد على طرق دمج العلاقات الدلالية بنسبة 2٪ لعلاقة مصطلح التبويب و نسبة 2٪ لعلاقة الارتباط و 3٪ لعلاقة holo member على التوالي. كان أفضل أداء لتصنيف النص عندما تكون علاقة holo member جزءًا من العلاقات المركبة حيث كانت النتيجة المتفوقة 81.2 للجمع بين علاقة الارتباط مع holo member بينما كانت النتيجة الأقل 78.6 لدمج علاقة الاسم الشامل مع مصطلح التبويب.

APPROVAL PAGE

The thesis of Musab Mustafa Hijazi has been approved by the following:

____
Akram M. Zeki M Khedher
Supervisor

____
Amelia Ismail
Co-Supervisor

____
Roslina Othman
Internal Examiner

____
Rosalina Abdul Salam
External Examiner

____
Mohamad Naqib Eishan Jan
Chairman

DECLARATION

I hereby declare that this thesis is the result of my investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Musab Mustafa Hijazi

Signature .... Date ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

ARABIC TEXT CLASSIFICATION BASED ON ARTIFICIAL BEE COLONY ALGORITHM AND SEMANTIC RELATIONS

I declare that the copyright holder of this thesis are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may only be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purpose.

3. The IIUM library will have the right to make, store in a retrieval system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Musab Mustafa Hijazi

……………………………… Signature

………………….. Date

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

ARABIC TEXT CLASSIFICATION BASED ON ARTIFICIAL BEE COLONY ALGORITHM AND SEMANTIC RELATIONS

I declare that the copyright holder of this thesis is International Islamic University Malaysia.

Musab Mustafa Hijazi

……………………………… ……………………..

Signature Date

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

ARABIC TEXT CLASSIFICATION BASED ON ARTIFICIAL BEE COLONY ALGORITHM AND SEMANTIC RELATIONS

I declare that the copyright holder of this thesis is Musab Mustafa Hijazi.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may only be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purpose.

3. The IIUM library will have the right to make, store in a retrieval system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Musab Mustafa Hijazi

……………………………… ……………………..

Signature Date

ACKNOWLEDGEMENTS

All glory is due to Allah, the Almighty, whose Grace and Mercies have been with me throughout my study. Although it has been tasking, His Mercies and Blessings on me ease the herculean task of completing this thesis.

To Prof. Dr. Akram and Dr. Amelia, whose enduring disposition, kindness, promptitude, thoroughness, and friendship have facilitated the successful completion of my work. who provided support, knowledge, and assistance to me at any time

To the great man who stood behind me supporting me throughout all my life, who taught me to be and how to be… my father, to whom taught me everything in this life and was sincere and endless supplications to the most precious of those in my heart… my mother, to those who shared with me every moment and all details of my life, to those who I do not like life without them, the title of joy and happiness, my brothers and sisters...

To everyone who taught me a letter... To all my friends who did not skimp on their support. To the adornment of the life and the happiness of days to the makers of the future and my hope in life to the flowers of my life, Jood, and Jana, and my smiling, my full of life, Mustafa. To my companion and life partner, my wife.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
AC	Associative Classification
ACC	Accuracy
ACM	Automatic Categorization Method
ACO	Ant Colony Optimization
AdaBoost.MH	multi-label boosting algorithm (extend for Adaptive Boosting)
AHP	Analytic Hierarchy Process
ANN	Artificial Neural Networks
ANNT	Artificial Neural Networks Training process
ARFF	Attribute Relation File Format
ARLStem	Arabic light stemmer
ARM	Associative Rule Mining
ASCII	American Standard Code for Information Interchange
ATC	Arabic Text Classification
AWN	Arabic WordNet
BABC	Binary Artificial Bee Colony
BALO	Binary Ant Lion Optimization
BAT	Binary Bat Algorithm
BDA	Binary Dragonfly Algorithm
BDA-SA	Binary Dragonfly Algorithm – Simulated Annealing
BGWO	Binary Gray Wolf Optimization
BiLSTM	Bidirectional Long Short-Term Memory
BNB	Bernoulli Naïve Bayesian
BNS	Bi-Normal Separation
BoCs	Bag of Concepts
BoWs	Bag of Words
BPNN	Back-Propagation Neural Networks
BPSO	Binary Particle Swarm Optimization

BR	Binary Relevance
BSO	Bee Swarm Optimization
CBA	Classification Based on Associations
CC	Correlation Coefficient
CDM	Class Discriminating Measure
CHI	Chi-Square
CN2	Clark & Nibblet (Induction Rules Algorithm)
CNB	Complement Naïve Bayes
CNN	Convolutional Neural Network
CP-1256	Code Page -1256
CT	Classification Trees
CTC	Compression-based Text Classification
DF	Document Frequency
DF_CF	Documents Frequency _ Category Frequency
DIA	Darmstadt Indexing Approach association factor
DL	Deep Learning
DMNB	Discriminative Multinomial Naïve Bayes
DT	Decision Tree
EMCAR	Expert Multiclass Classification based on Association Rules
ERR	Error Rate
EST	Educated Text Stemmer
FA	Field Association
FACA	Fast Associative Classification Algorithm
FAFS	Firefly Algorithm Feature Selection
FN	False Negative
FP	False Positive
FRAM	Frequency Ratio Accumulation Method
FS	Feature Selection
GA	Genetic Algorithm
GABC	Global Artificial Bee Colony
GBDT	Gradient Boosting Decision Tree
GI	Gini Index

GNB	Gaussian Naïve Bayes
GR	Gain Ratio
GRU	Gated recurrent unit
GSS	Galavotti-Sebastiani-Simi Coefficient
HANGRU	Hierarchical Attention Network- Gated Recurrent Unit Deep Learning Model
HMM	Hidden Markov Model
HPABC	Hybrid Particle-move Artificial Bee Colony algorithm
ICF	Inverse Class Frequency
IDF	Inverse Document Frequency
IG	Information Gain
ISRI	Information Science Research Institute
ISVM	Improved Support Vector Machine
ITF	Inverse Term Frequency
KNB	Kernel Naïve Bayes
KNN	K-Nearest Neighbors
LC	Label Combination method
LDA	Linear Discriminant Analysis
LogTF	Log Term Frequency
LOOCV	Leave One Out Cross-Validation
LOPS	List of Pertinent Synsets
LOPW	List of Pertinent Words
LR	Logistic Regression
LSI	Latent Semantic Indexing
LSTM	Long Short Term Memory
LTC	Lookup Table Convolution
MBNB	Multi-variant Bernoulli Naïve Bayes
MCAR	Multi-Class Association Rule
ME	Maximum Entropy
MI	Mutual Information
MLP	Multilayer Perceptron
MLP-NN	Multilayer Perceptron Neural Network
MLR	Multinomial Logistic Regression

MNB	Multinomial Naïve Bayes
MR	Modification Rate
MW	Maximum Weight
NB	Naïve Bayes
NBM	Naïve Bayes Multinomial
NGL	Ng-Goh-Low Coefficient
NLTK	Natural Language Toolkit
NN	Neural Network
OCATC	Optimal Configuration Determination for Arabic Text Classification
OR	Odds Ratio
P	Precision
PART	Partial Decision Tree Algorithm (developed version of C4.5)
PCA	Principle Component Analysis
PNNs	Polynomial Neural Networks
PSO	Particle Swarm Optimization
R	Recall
RBF	Radial Basis Function
RCV1	Reuters Corpus Volume 1
Relief F	an extension of the original Relief algorithm
REP	Reduced Error Pruning
RF	Random Forest
RFBoost	accelerated version of AdaBoost.MH
RIPPER	Repeated Incremental Pruning to Produce Error Reduction
RS	Relevancy Score
SA	Simulated Annealing
SACM	Semi-Automatic Categorization Method
SF-MW	Semantic Fusion- Multiple Words
SGD	Stochastic Gradient Descent Algorithm
SVM	Support Vector Machines
TC	Text Classification, Text Categorization
TF	Term Frequency
TFICF	Term Frequency Inverse Class Frequency

TFIDF	Term Frequency-Inverse Document Frequency
TN	True Negative
TP	True Positive
TR	Triggers Classifier
TS	Term Strength
UTF	Unicode Transformation Format
WC	Word Count
WIDF	Weighted Inverse Document Frequency
WLLR	Weighted Log-Likelihood Ratio
WSD	Word Sense Disambiguation
XGBoost	an ensemble technique of decision trees and a variant of gradient boosting algorithm.

CHAPTER ONE

INTRODUCTION

1.1. BACKGROUND OF STUDY

Documentation is the best method to illustrate knowledge, which implies that the substantial repositories of information are documents (Khorsheed & Al-Thubaity, 2013). Due to the rapid expansion of the internet, there is a massive growth in the number of electronic documents, which require flexible and effective ways to access, arrange, and extract useful information, such as text classification and text clustering (Riyad Al-Shalabi & Obeidat, 2008; Khorsheed & Al-Thubaity, 2013). Text classification is the process of grouping or categorizing documents into pre-defined groups or classes based on pre-defined criteria. (Rasha Elhassan & Ahmed, 2015b; Khreisat, 2006). Text classification (TC) has been utilized in several applications including document organization, text filtering, document automated indexing, spam filtering, and Disambiguation of words meaning or sense (Riyad Al-Shalabi & Obeidat, 2008).

Information overburden is a raised issue of data preparation and collection in research work. This implies a waste of time in the data analysis process. So, having accurate and efficient low-dimensional data from a high-dimensional one is needed. Analysis of enormously high-dimensional data, by removing unnecessary ones, is a substantial process in data mining called dimension reduction. Feature selection (FS) is one of the dimension reduction processes (H. K. H. Chantar, 2013; Prasartvit et al., 2012; Shunmugapriya & Kanmani, 2017). Feature selection is used in text classification to enhance computational efficiency and classification accuracy by eliminating redundant or unnecessary words (features) and holding only the features that have pertinent information to simplify the classification process. Typically, there is a huge list of features, and many of them are not useful for TC, so robust FS methods are required to have accurate and efficient low-dimensional data from high-dimensional ones. Feature selection has two types of methods: wrapper and filter method. In the wrapper method, features are picked out or filtered