

AN INTELLIGENT MODEL FOR AGE AND GENDER PREDICTION FROM ARABIC  
TWITTER USER'S POSTS.

BY:

FAHAD AL MEKHLAFI

A THESIS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF:

MASTER OF SCIENCE IN INFORMATION SECURITY

TO:

INFORMATION SECURITY DEPARTMENT

THE COLLEGE OF COMPUTER AND INFORMATION SECURITY  
NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

2017G-1438H

SUPERVISED BY:

DR. HUSSEIN Y. MANSOUR

NAIF ARAB UNIVERSITY FOR SECURITY SCIENCES

## Table of Contents

Dedication .....	iv
Acknowledgement .....	v
List of Figures .....	viii
List of Tables .....	ix
List of Acronyms .....	x
1 Introduction.....	1
1.1 Introduction.....	1
1.2 Research Aims, Scope and Objectives.....	2
1.3 Motivation.....	2
1.4 Thesis Contributions .....	5
1.4.1 Enhancing the tweets pre-process phase seeking accuracy enhancement .....	5
1.4.2 designing a model capable of dealing with Arabic dialects.....	5
1.4.3 Propose new technique to deal with multi label classification problem.....	5
1.4.4 Extensive literature review on classification .....	6
1.4.5 Validated dataset.....	6
1.5 Research methodology.....	6
1.6 General Structural Design.....	8
1.7 Thesis outline .....	9
2 Background and Literature Review .....	10
2.1 Introduction.....	10
2.2 Data Mining .....	11
2.3 Text Mining .....	20
2.4 Common classification techniques in datamining: .....	20
2.4.1 Statistical procedure based approach:.....	20
2.4.2 Machine learning based approach:.....	21
2.4.3 Neural networks .....	21
2.5 Classification Algorithms in datamining: .....	22
2.5.1 Binary Classification:.....	22
2.5.2 Multiclass Classification.....	22
2.5.3 Multilabel Classification.....	23
2.5.4 Multiclass-Multilabel Classification.....	26
2.6 Common Algorithms for Multilabel Classification .....	26
2.6.1 The Data Transformation Approach .....	27
2.6.1.1 Binary Relevance (BR).....	28
2.6.1.2 Label PowerSet (LP).....	30
2.6.2 The Method Adaptation Approach .....	31
2.6.2.1 Tree-Based Methods .....	31
2.6.2.2 Neuronal Network-Based Methods.....	32
2.6.2.3 Vector Support Machine-Based Methods.....	32
2.6.2.4 Instance-Based Methods:.....	32
2.7 Common text classification Algorithms .....	33
2.7.1 Decision trees.....	33
2.7.2 Pattern (Rule)-based Classifiers.....	33
2.7.3 SVM Classifiers .....	34
2.7.4 K Nearest Neighbors k-nearest neighbors .....	34
2.7.5 Neural networks classifiers.....	34
2.8 Text classification domains: .....	34

2.9	Author Profiling.....	35
2.10	Extraction and Selection Feature: .....	36
2.10.1.1	Feature Extraction.....	36
1.	Content-Based Features .....	37
2.	Statistical -Based Features .....	38
2.10.1.2	Feature Generation:.....	40
2.10.1.3	Feature Selection.....	42
2.10.1.3.1	The feature selection techniques:.....	43
2.11	Related Works:.....	44
2.11.1	Arabic text related work: .....	46
2.11.2	English text Related Work:.....	50
2.11.3	Summary .....	54
3	The Proposed Preprocessing Data and Extract Features Model.....	55
3.1	Introduction:.....	55
3.2	Data Collection .....	56
3.3	Preprocessing Tweets: .....	58
3.3.1	Processing Tweets:.....	59
3.3.1.1	Emoji Processing: .....	64
3.3.1.2	Emoji Process experimental test:.....	67
3.3.2	Impact of Pre-Processing on classification accuracy:.....	69
3.3.2.1	Extract feature Experimental .....	75
3.4	Summary .....	79
4	The Adaptation of the Proposed Model to Arabic twitter profiles: A Case Study.....	80
4.1	Introduction.....	80
4.2	Empirical study and experimentations for age and gender prediction:.....	82
4.2.1	Data Collection .....	82
4.2.2	Text preprocessing:.....	83
4.2.3	Feature extraction: .....	83
4.2.4	Training and Evaluating the Classifier .....	85
4.2.5	Evaluation Measures for the prediction:.....	88
4.2.6	Prediction Results: .....	90
4.2.7	Summery .....	92
5	Conclusions and Future work .....	93
5.1	Thesis Contribution.....	93
5.2	Limitations of the Current Work.....	94
5.3	Future Work and research trends .....	94
5.4	Conclusion .....	95
6	Bibliography .....	96

## List of Figures

Figure 1.1 General Structural Design .....	8
Figure 2.1. DM Methods attending to the available data nature [4].....	12
Figure 2.2 . Decision boundaries resulting from some of the best-known learning models[4] .....	14
Figure 2.3. Iris species categorization is a classical multiclass classification problem[4]. ....	23
Figure 2.4. Image labeling is an usual multilabel classification task[4].....	25
Figure 2.5. Categorization of representative multi-label learning algorithms being reviewed	27
Figure 2.6. Binary Relevance transformation diagram[4]. .....	28
Figure 2.7. Label Powerset transformation diagram. [4].....	30
Figure 3.1. Collect User Profile Tweets Model .....	57
Figure 3.2. Tweets Preprocessing Model.....	59
Figure 3.3 Process tweets text file .....	60
Figure 3.4. Sub-Process Process Documents Image.....	61
Figure 3.5 Arabic light stemming algorithm steps.....	63
Figure 3.6 Tweets Emoji Dictionary Processing Model.....	65
Figure 3.7 Sample of Twitter Emoji (Twemoji) Dictionary .....	66
Figure 3.8 No. of Attributes of Original, Emoji Dictionary and Removing Emoji .....	68
Figure 3.9 No. of attributes of Original, Emoji Dictionary and Removing Emoji .....	68
Figure 3.10 Accuracy of Original, Emoji Dictionary and Removing Emoji .....	69
Figure 3.11. Original tweets file before processing by Tweets Preprocessing Model .....	70
Figure 3.12. New tweets file after processing by Tweets Preprocessing Model .....	70
Figure 3.13 Examples of Processing Tweets ,The average=182952.1 .....	73
Figure 3.14 Attributes of Processing Tweets ,The average=41670.13 .....	73
Figure 3.15 Accuracy of Processing Tweets The average =53% .....	74
Figure 3.16 Time of Processing Tweets The average =6.4 .....	74
Figure 4.1. Build Dataset Model.....	84
Figure 4.2. Classifier Model a.....	85
Figure 4.3. Classifier Model b(Training).....	86
Figure 4.4. Classifier Model C(Testing).....	86
Figure 4.5. K-fold cross validation .....	87
Figure 4.6 Prediction Accuracy Naive Byas vs DL Nuoral Networks .....	91

## List of Tables

Table 2.1 Classification problems attending to the output to be predicted [4].....	18
Table 3.1 Sample of modifying title of Emojis.....	66
Table 3.2 Accuracy, Example and Attributes Result Using Emoji Dictionary .....	67
Table 3.3 Processing Tweets Performance and Accuracy.....	72
Table 3.4 Bag of Words .....	75
Table 3.5 Total words in each document.....	76
Table 3.6 Example Of Term Frequency In Twitter Dataset For User Profile .....	76
Table 3.7 Example of Inverse Document Frequency (IDF) In Twitter Dataset For User Profile.....	76
Table 3.8 Example of Term Frequency -Inverse Document Frequency (TF-IDF) In Twitter Dataset For User Profile.....	77
Table 3.9 Example of Binary Term Occurrence of Twitter Dataset For User Profile.....	78
Table 3.10 Example of Trigram of Twitter Dataset For User Profile.....	78
Table 4.1 Tweets of Profile Samples .....	82
Table 4.2 Accuracy ,Precision and Recall Using Naïve Bayes .....	91
Table 4.3 Accuracy ,Precision and Recall Using Deep Learning Neural Networks .....	92

## List of Acronyms

Acronyms	Details
ANN	Artificial Neural Networks.
URL	Uniform Resource Locator.
KDD	Knowledge Discovery and Data Mining
AI	Artificial Intelligent.
ML	Machine Learning.
DL	Deep Learning.
DM	Data Mining.
KNN	K Nearest Neighbors.
SVM	Support Vector Machines
SVN	Support Vector Network.
LIBSVM	Library for Support Vector Machine.
OVA	One-Vs-All.
OVO	One-Vs-One.
MDs	Multilabel Datasets.
MC	Multilabel Classification.
MLP	Multilabel Prediction.
DT	Data Transformations.
BR	Binary Relevance.
LP	Label PowerSet.
RPC	Ranking by Pairwise Comparison.
CLR	Calibrated Label Ranking.
MEKA	Multi-label Extension to WEKA.

Acronyms	Details
WEKA	Waikato Environment for Knowledge Analysis.
LC	Label Calibration.
RT	Ranking and Threshold.
TC	Text Classification.
NB	Naïve Bays
ML	Multi Label
MFCC	Mel-Frequency Cepstral Coefficients.
SFM	Spectral Flatness Measure.
RBFKernel	Radial Basis Function Kernel.
LIBINAR	Library for Large Linear Classification
TF-IDF	Term Frequency-Inverse Document Frequency
HTML	Hyper Text Markup Language
Ws	Slang Word.
LIWC	Words Linguistic Inquiry and Word Count dictionary.
BoW	Bag of words.
TF	Term Frequency.
TO	Term Occurrences.
BTO	Binary Term Occurrences.
VSM	Vector Space Model.
ARI	Automated Readability Index.
CLI	The Coleman-Liau Index.
RIX	The Rix Readability Index.
NLP	Natural Languages Processing.

<b>Acronyms</b>	<b>Details</b>
LSI	Latent Semantic Indexing.
LPI	Locality Preserving Indexing.