

AMERICAN UNIVERSITY OF BEIRUT

Emotion Recognition From Text and
Physiological Data

by

Ramy Anwar Awwad

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Electrical and Computer Engineering
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon

May 2014

AMERICAN UNIVERSITY OF BEIRUT

Emotion Recognition From Text and Physiological Data

by
Ramy Anwar Awwad

Approved by:

Dr. Hazem Hajj, Associate Professor
Electrical and Computer Engineering

Advisor



Dr. Zaher Dawy, Associate Professor
Electrical and Computer Engineering

Member of Committee



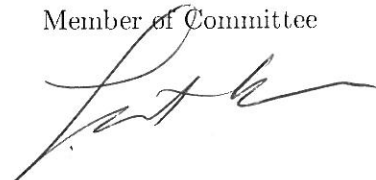
Dr. Wassim El Hajj, Assistant Professor
Computer Science

Member of Committee



Dr. Fatima Al-Jamil, Assistant Professor
Psychology

Member of Committee



Date of thesis defense: May 5th, 2014

Acknowledgements

First, I thank God for his blessings and the benefits he gave me in order to accomplish this work, and without whom nothing was possible. Secondly, I offer my sincerest gratitude to my supervisor, Prof. Hazem Hajj who has supported me during my graduate studies at AUB. In addition, I want to thank all the committee members Prof. Wassim El Hajj, Prof. Fatima Al- Jamil, and Prof. Zaher Dawy for their help and support. I also want to thank Intel for their support and guidance, in particular, Director of MER program Dr. Tawfik Arabi, Director of Context Aware Sensing Program, Mrs. Lama Nachman, and Physiology Project Mentor Dr. Jennifer Healey. I also want to thank KACST for their support and guidance in particular Dr. Abdulfattah Obeid and Dr. Mohammad Ben Saleh Finally, I would like to thank my family and friends who supported me all the time especially during periods of difficulty.

An Abstract of the Thesis of

Ramy Anwar Awwad for Master of Engineering
Major: Software Systems

Title:Emotion Recognition From Text and Physiological Data

Ability to feel emotions is known to be an intrinsic property of human beings. At every moment of our life, we unconsciously respond emotionally to everything that happens around us. Emotions are normally triggered by a person's environment and context, and then expressed through physiological changes and physical reactions. In this thesis we explore two modalities for emotion recognition: Text and Physiological. Among possible choices of context, text is the most common source of personal data. Physiological measurements, on the other hand, should present one of the common ways of showing emotional reactions. For emotion recognition from text, we propose a new method for optimal feature selection. We call this method MFX (the most frequent and discriminative features across). For physiological data, we conduct a study for improving accuracy of Ground truth data, and an evaluation for ranking of most relevant Physiological features. Experiment results with MFX show superiority for text classification with benchmark RCV1 and Reuters Data, and 85% accuracy for emotion recognition from text. On the other hand, experiments with Physiological data, showed that raters' assessments provide more accurate ground truth data and higher recognition accuracies. Furthermore, the results of ranking physiological features indicate a correlation between certain features and certain aspects of emotions. For example, features extracted from Galvanic Skin Response (GSR) correlated more with Valence.

Contents

1	Introduction	1
2	Emotion Recognition from Text	3
2.1	Feature Selection for Text Categorization	3
2.1.1	Statistical-Based Approaches for Feature Selection . . .	3
2.1.2	Semantic-Based Approaches for Feature Selection . . .	5
2.1.3	Arabic Feature Selection	5
2.2	Emotion Classification	6
2.3	Proposed Feature Selection Methodology	7
2.3.1	Optimization Approach to Feature Selection	7
2.3.2	Feature Extraction and Selection with MFX	8
3	Results and Analysis	14
3.1	News Classification	14
3.1.1	MFX Evaluation on Small to Medium Sized Arabic Documents	14
3.1.2	MFX Evaluation on Larger Sized Arabic Documents . . .	16
3.1.3	Benchmark evaluation with RCV1 dataset	18
3.1.4	MFX Evaluation on Reuters-21578 dataset with Conventional Methods	18
3.1.5	MFX Evaluation on Reuters-21578 dataset with Modern Methods	20
3.2	Emotion Classification	20
3.2.1	Dataset used for Emotion Recognition from Text . . .	21
3.2.2	Experiment results for Emotion Recognition	22
3.3	Conclusion	26
4	Emotion Recognition from Physiological Data in Natural Settings	27
4.1	Literature Review	27

4.2	Proposed Approach for Emotion Recognition from Physiological Data in Natural Settings	30
4.2.1	Method for Ground truth Collection in Natural Setting	31
4.2.2	Inter-Raters Agreement Calculation	32
4.3	Evaluation and Ranking of Physiological Features for Emotion Recognition	33
4.4	Experiment for Creating Ground Truth Data	34
4.4.1	Data Labeling	35
4.4.2	Physiological features extractions	36
4.4.3	Data Cleaning	37
4.4.4	Training Data creation	38
4.5	Observations	38
4.5.1	Valence Evaluation	38
4.5.2	Arousal Evaluation	39
4.5.3	Control Evaluation	41
4.6	Conclusion	42
5	Conclusion	43

List of Figures

1.1	Flow Chart of Emotion Recognition from Text and Physiological Data	1
2.1	Term Summary matrix with terms as rows and categories as columns; used to store the weight of every term in every category	9
3.1	Classification accuracy when document sizes vary from low to medium and when categories are close and distinct	15
3.2	MFX comparison to other feature selection algorithms when applied to relatively large Arabic documents (6kB)	17
3.3	MFX comparison with SIPs when applied to relatively large Arabic documents	18
3.4	MFX comparison with conventional methods when applied to RCV1 dataset	19
3.5	MFX comparison to conventional feature selection methods for English text	19
3.6	MFX comparison with SIPs when applied to English documents	20
3.7	MFX comparison to modern classification methods for English text	21
3.8	MFX Performance on Amman Dataset	24
3.9	MFX Performance on UIUC Children’s Story Corpus Dataset	25
3.10	relation between term and the associated emotional class	25
3.11	MFX Performance on ISEAR Dataset	26
4.1	Data Collection Model	32
4.2	Proposed Template for Raters	32
4.3	Participant Event	34
4.4	Preprocessing Phase	35
4.5	Data Reduction	38
4.6	Participants Distribution	38
4.7	Evaluation of Valence Model	39
4.8	Evaluation of Arousal Model	40

4.9	Evaluation of Control Model	41
-----	---------------------------------------	----

List of Tables

3.1	Dataset from Al- Jazeera newspaper and OSCA dataset	17
3.2	ISEAR summary dataset	21
3.3	Amman summary dataset	22
3.4	UIUC Children’s Story Corpus	23
3.5	List of features selection based on δ	23
3.6	List of rejected features based on δ	24
4.1	Summary of Sensor and Corresponding Bio-Signals	28
4.2	Survey of Methods for Emotion Recognition using Physiological Signal	29
4.3	Alpha Measures for Different Emotional Classes	33
4.4	Summary of the Classification of the Instances	36
4.5	Features Extracted from the Physiological Sensors	37
4.6	Features Selection by Applying Forward and Backward Elimination for Valence	39
4.7	Features Selection by Applying Forward and Backward Elimination for Arousal	40
4.8	Features Selection by Applying Forward and Backward Elimination for Control	42

Chapter 1

Introduction

Researchers have tried to build systems for Emotions recognition using different types of data sources, including physiological sensors, gesture, facial expressions, and textual expressions [1]. We propose to use physiological data since it presents one of the most reliable source for emotion recognition. We also propose the use of textual data since it is the most commonly used of data associated with humans. A flowchart of emotion recognition from text and physiological data is presented in Figure 1.1

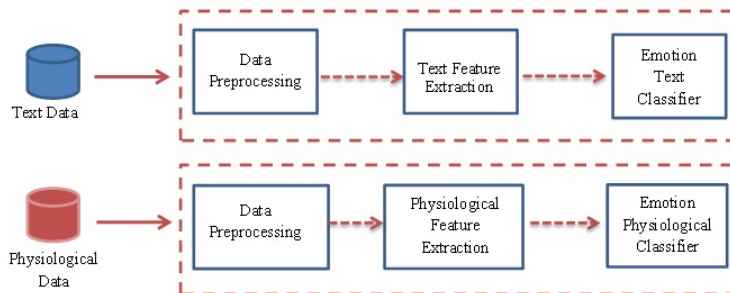


Figure 1.1: Flow Chart of Emotion Recognition from Text and Physiological Data

For emotion recognition from text, researchers have focused their interest on extracting emotions from texts. The main features extraction methodologies are based on emotion generation rules [2], emotion keywords [3], Learning-based Methods [4], emotion ontology [5]. However in several cases, emotions are not expressed by using words with an affective meaning (e.g. happy or sad), but by illustrating real life situations, where readers are able to relate them to a specific emotion [6]. It is relatively easy for a human being to

understand the emotions by processing the text. But, when it comes to intelligent devices such as: smart phones and computers, the process is more complex due to the difficulty in automatic inference of meaning from text. We propose a new method for emotion recognition from text with optimal choice for number of features needed. The problem of feature selection is formulated as an optimization problem.

For Physiological data, the use of this source of modality for emotions recognition has gained a lot of interest. Several researchers have focused their studies on feature extraction and feature reduction from physiological sensors [7,8]. They identified basic statistics of the physiological measurements as key features in emotion classification. These features are also be used in our study. Furthermore, most of the past research on emotion recognition was conducted in lab settings [9], and very few researchers attempted to develop models in natural settings [10]. The accuracy of the models developed in natural settings remains a challenge. Our objective in this regard is to develop an accurate emotion recognition system from physiological data in natural settings. We build our research on a clinical study that was conducted with 15 participants, where physiological data was gathered during normal day to day activity, along with self emotion annotations. In this proposal, to use the clinical study to generate accurate ground truth emotion data. We use independent expert evaluators to improve the accuracy of the Ground truth. The results of the in-dependent assessments are combined to provide the accurate training data for building the emotion recognition model.

The outline of the thesis is as follows. In chapter 2, we present the proposed method for emotion recognition from text. In chapter 3, we present the proposed evaluation for emotion recognition from physiological data. Chapter 4 contains the summary and future work.

Chapter 2

Emotion Recognition from Text

In this chapter, we start in section 2.1 with a review of the work done on feature selection in text classification for English and Arabic text. In section 2.2 we review the recent work on emotion recognition from text. In section 2.3, we present our proposed methodology for optimal features selection from text. In section 2.4 and 2.5 describe the experiments conducted on news and emotions-annotated text respectively. Finally, we present a summary of the work done and highlight the remaining open challenges as a future work.

2.1 Feature Selection for Text Categorization

In this section, we present a summary of the work related to feature selection for text categorization. We divide the section into three subsections based on the approach used for feature selection. We start by reviewing statistical-based approaches, followed by semantic-based approaches, and end by discussing the approaches specific for the Arabic language.

2.1.1 Statistical-Based Approaches for Feature Selection

Emotion recognition from text can be considered as a text categorization problem, where the categories are one of the emotions such as happy, sad, and disgust. As a result, research from text categorization can be directly applicable for this thesis. Several text feature selection methodologies have been used in the past years, with particular focus on considering words or word combinations as terms whose weights and counts represent features for text categorization. Some of the methods give higher weights to terms that are more prevalent or significant across documents. The simplest such

technique is Document Frequency (DF) [11]. Other techniques only measure significance of the terms within a document such as Term Frequency (TF), Darmstadt Indexing Approach (DIA) association factor, and Ambiguity measure (AM) [12–14]. Both feature selection techniques, the ones that consider terms within documents and those that consider terms across documents, provide acceptable accuracy in text categorization. Consequently, researchers tried to find a measure that comprehends the importance of the terms within and across documents [15–21]. TF-IDF [20] was the first method that combined the characteristics of the intra-category and the inter-category measures. Later, chi-square, mutual information (MI) [11], and Information Gain (IG) [15] were widely used. Chi-square measures the independence between the terms and a specific category. IG gives a measure of relevance between features and class labels. In [11], it was reported that both Chi-Square and IG are effective feature selection algorithms. Consequently, IG became popular method for feature selection. However, the selection of the top features remains arbitrary using a predefined threshold. Many works, including ours, improve on the mentioned approaches for the purpose of eventually getting higher classification accuracy. In [17], an improved Gini Index methodology was introduced to measure the purity of a term towards the category. In [19], a method called Orthogonal Centroid Feature Selection (OCFS) was introduced. Similar to IG, OCFS selects the best set of features by solving an objective function based on the orthogonal centroid algorithm. In [21], Yang et.al discuss a new technique called Comprehensively Measure Feature Selection (CMFS), which is a combination of simultaneously selecting and weighing the terms Their experiments showed that CMFS was superior to Chi-Square, OCFS, IG, DF, and DIA. In [16], Odd ratio is presented as an evaluation function based on the ratio of the probability of the term belonging to a given a category over the probability of the term belonging to other categories. In [18], the feature selection methodology is divided into two stages. In the first stage, information gain is used as ranking criteria to determine the importance of the features. In the second stage and towards dimension reduction, Principal Component Analysis (PCA) and Genetic Algorithm (GA) are applied to the terms with the highest importance ranked by IG. In [22], a feature selection method called Statistically Improbable Phrase Selection (SIPs) is suggested. SIPs is used by Amazon and rotates around identifying key-words of a given book in the search engine. The algorithm searches for the most distinctive phrases in the text. Phrases that occur frequently are considered the SIPs for a particular book. However, all the methods try to select the best set of features based on a best-guess threshold depending on the method. As a result, they all lack a systematic approach of finding an optimum threshold for selecting the best features. In

our approach, we provide a mathematical formulation for determining the set of terms that best differentiates the different categories.

2.1.2 Semantic-Based Approaches for Feature Selection

Other researchers proposed non-term based methods for feature selection such as distributional clustering [23] [15]. The objective was to build a classifier for each category. In the training phase for each category the authors selected from documents a number of terms labeled with positive and negative sentiments. The documents were then represented by the set of positive and negative terms. L-square was used to generate the rules of classification based on learning logic.

2.1.3 Arabic Feature Selection

Specific to the Arabic language, researchers focused on the morphological part, by developing morphological stemmers that aim to enhance feature selection methods [24–26]. In [24], the authors applied stemming process by adopting the work of AL-Shalabi [27] for root extraction. The method suggests assigning weights for a word’s letters multiplied by the letter’s position inside the word. The stemmed words were stored in feature vectors and performance analysis was conducted using three classifiers: Naïve Bayes (NB), KNN, and distance based classifiers. The results showed that NB outperformed the other classifiers. In [25], the authors adapted the same stemming process suggested in [26] producing a low accuracy of 63 percent when NB was used. The work in [24], just mentioned above, also improves on the work of [25, 26] by using TF and DF in addition to the proposed stemming process. When tested on the same datasets using decision trees as classifier, the proposed approach outperformed the approaches of [25, 26] by achieving a 6 percent improvement on the classification accuracy. In addition to achieving improved accuracy, the work in [24] showed that the stemming process has some limitations and it is not the best feature selection approach. Other techniques have been proposed to enhance the feature selection methodologies as described in [28–30]. In [29], the Chi-Square technique was implemented for Arabic. The best reported accuracy was 88 percent. In [28], the authors proposed a text classification method based on N-grams. The major advantage of this approach over many others is that it is based on selecting a bunch of letters to represent text making it language independent. On the other hand, the probability of having similar terms in the various categories will increase specially for 2-grams and 3-grams. Thus, adding more complexity to the classification process. In [30], the authors presented a fea-

ture selection approach based on ant colony optimization. Their algorithm is divided into several steps. It starts by selecting the features based on TF-IDF followed by generating the feature set. The algorithm then uses the ant colony optimization algorithm in order to select the best subset of features. However, it is difficult to find the importance of the terms since the ranking value of these features varies among the negative and positive categories. Our approach, MFX, is a statistical-based approach that considers all documents from the same category as one extended document, and chooses the most discriminative terms that are frequent and common across all documents of the same category, but rarely present in other categories. MFX is language independent and backed up with a mathematical formulation that finds the optimal number of features that guarantees accurate text categorization. We discuss next the details of MFX and the accompanying mathematical formulation.

2.2 Emotion Classification

Most of the work done on emotion recognition from text is related to feature extraction methodologies which can be divided into three model representations [31]. The first one consists of corpus based model such as in [5,32,33]. This technique includes the use of emotion lexicon that contains words with corresponding scores representing how well they correlate with different emotions. The weights of the words are then used as features for classification. The second set of approaches use statistical measures for deriving the weights of the terms in association with emotions [34,35] such as the use of Term frequency (TF) [36]. In [37], they provide a comparative study among different statistical technique such as: unigram, and trigram. In [38] they used a technique similar to the Document Frequency (DF). They showed that automatically selecting feature performed better than the manual selection which is based on the wordNet-Affect [39]. The last category is uses knowledge base to determine semantics in sentences that trigger certain emotions. These techniques result in the generation of emotion rules [2]. In [40] they provide an Ontology Knowledge model to extract the semantic features from sentences. Others like [41] work on a hybrid approach which rely on features extracted via statistical measures and semantic features by using emotional lexicon database such as WordNet-Affect. The statistical based algorithms are the most widely used in emotion detection because they are easily adaptable and language independent. Our proposed methodology is designed to work with either statistical approaches or corpus-based approaches.

2.3 Proposed Feature Selection Methodology

We start this section by describing our proposed feature selection approach by an optimization formulation with the objective of maximizing classification accuracy that accurately clusters the features of the different categories. Then, we discuss the new discriminative feature that we called Most Frequent Common Words across Documents (MFX).

2.3.1 Optimization Approach to Feature Selection

We address now the problem of finding an optimal threshold to select the best subset of features in a given category. Hence, the optimal threshold guarantees achieving the highest classification accuracy. Traditionally, the value of a threshold is chosen by trial and error, where many thresholds are tested and the one producing the highest classification accuracy is chosen. On the other hand, we formulate the problem mathematically as an optimization problem that aims at finding the optimal threshold that minimizes the distance among words within a category (cluster) while at the same time maximizes the distance between different categories (cluster). The idea is to pick the features that will eventually yield the best classification results. The problem of obtaining the features that best represent each category can be formulated as follows:

$$\underset{\delta, \tau_j, 1 \leq j \leq c}{\operatorname{argmax}} \left(\sum_{j=1}^c \left(\sum_{k=1, k \neq j}^c \frac{D_{jk}}{|c-1|} - d_j \right) \right) \quad (2.1)$$

Subject to the following constraints:

- c : Total number of categories
- τ_j : Decision variable giving threshold to select the best features in a category j
- δ : Decision variable giving selectivity threshold for discrimination between different categories. Terms that have high discriminative capability across categories are selected if their selectivity defined by the following equation is above the selectivity threshold δ :

$$\left(\frac{\max(\text{termweights}) - \text{median}(\text{termweights})}{\max(\text{termweights})} \right) \geq \delta \quad (2.2)$$

- y_i : Number of terms chosen from category j based on threshold τ_j

$$y_i = \tau_j * N_j \quad \forall j \quad (2.3)$$

- N_j : Total number of terms in category j
- v_j : Center of the cluster of category j

$$v_j = \frac{1}{|doc|_j} \sum_{i=1}^{|doc|_j} u_{ij} \quad \forall j \quad (2.4)$$

- u_{ij} : is a vector of weights for the terms in doc i and category j . The weights can be computed by one of the conventional measures such as TF or TF-IDF. $|doc|_j$ is the number of documents in category j . $u_{ij} = [z_{1ij}, \dots, z_{y_j ij}]$
- $z_{y_j, ij}$: Weight of term y_j in doc i in category j . $z_{y_j, ij}$ calculation is further presented in lines 1-8 of the pseudo code in Algorithm 1
- d_j : Intra-category distance between the y_j terms terms that are left in category j after applying the threshold τ_j . Note that d_j represents the strength of the correlation between the terms of a certain cluster.

$$d_j = \sum_{i=1}^{|doc|_j} \sum_{k=1}^{\tau_j * N_j} ||u_{ik} - v_k||^2 \quad \forall j \quad (2.5)$$

- D_{jk} : Distance between documents with category j and category k based on calculating the distance between the centers of the categories

$$D_{jk} = ||v_j - v_k||^2 \quad \forall_{j,k} j \neq k \quad (2.6)$$

The objective function finds the optimum thresholds that maximize the distances between the clusters or categories and minimize the distances between terms inside the clusters.

2.3.2 Feature Extraction and Selection with MFX

For feature weights, our proposed approach uses features weights with some resemblance to TF-IDF and SIPs in the sense that it obtains word statistics across all documents. Other weights can also be used. The major contribution is that we propose to eliminate non-discriminative terms that are

frequent and common across all documents in multiple categories (line 11-18 in Algorithm 1). We combine our proposed feature choice with the optimization approach (section 2.3.1) to derive an optimal number of features, rather than a threshold based on predefined experiments. Given c categories and d documents, MFX selects the best features to represent every category leading to higher accuracy in the document categorization i.e. which category the document belongs to. In addition, to selecting the best features for each category, MFX also helps in selecting the discriminant terms for each category and thus increasing the classifier accuracy. In what follows we use *word* and *term* interchangeably.

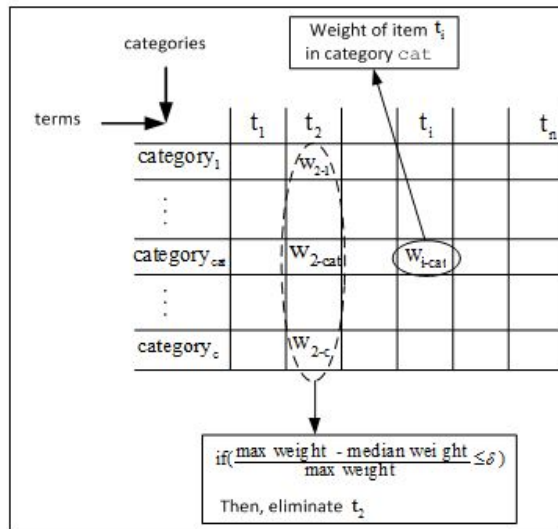


Figure 2.1: Term Summary matrix with terms as rows and categories as columns; used to store the weight of every term in every category

The pseudo code for the MFX algorithm is presented in Algorithm 1. cw_{cat} is an array of common words in category cat , w_{i-cat} is the weight of each term t_i in category cat , and, $termSummary$ is a matrix with terms as rows and categories as columns used to store the w_{i-cat} values Figure 2.1. MFX is divided into four major steps:

- **Step 1: Feature weights:** For this step, we propose to use normalized Term Frequency (TF) - Line (1) through Line (10): For each category, we get all the common words across all documents, count them, and calculate the normalized weight for each word based on its frequency, the sizes of the documents, and the number of documents corresponding to a particular category (line 8). Line (9) stores the weights of the

terms in their corresponding positions in the matrix *termSummary*. The resultant weighting scheme gives a high weight to the most frequent terms, allowing the method to be more accurate when selecting discriminating terms. This process is repeated for all categories. After the completion of this step (lines 1-10), matrix *termSummary* is fully populated; every column corresponds to a term along with its weight across categories.

- **Step 2: Eliminating non-discriminating terms Line (12):** This step aims at only keeping the discriminating terms, while eliminating the non-discriminating ones. For this paper, a simple discriminating measure is used. The logic is as follows: if the number of occurrences of a term (term weight) is high in one category and low in other categories, the term should be considered a discriminating feature and should be considered in the classification. We capture this idea in line (12) of Algorithm 1, where for every feature, we find the maximum weight and the median across the existing categories. We then apply the formula in line (11-iii) to eliminate terms that are not discriminating based on the selectivity threshold \hat{t} that is varied between 0 and 1.
- **Step 3: Optimal Threshold Selection and Deriving Best Features for each Category Lines (11-18):** This step follows the optimization described in the previous section (2.3.1). For a given selectivity threshold δ , and for each category i.e. for every column in the *termSummary* matrix, the method selects the most relevant terms, the terms with the highest weights. The loop in the algorithm (lines 11-18) determines the best selectivity threshold and the best percentage of terms to pick as most relevant. As described in section (2.3.1), the threshold selection is based on maximizing the accuracy of classification, which is represented by the thresholds leading to the highest separation in multi-dimensional space between the groups of documents representing different categories. Hence, the selected words, stored in array *cw*, best represent each category because (1) statistically, they are the most frequent across the documents in a specific category and (2) quantitatively, their cardinality is optimal since the number of features is chosen based on a deterministic optimized mathematical formulation.
- **Step 4: Documents-Terms Feature Matrix - Line (19) through Line - (24):** The last step in the algorithm is to scan the dataset again based on the terms in array *cw* i.e. terms selected after running lines (11) through (18). For every document in every category, we record the feature weights for every word belonging to *cw*, and assign it to the

proper feature term. Every document is annotated with the category class it comes from. For example, if the word "soccer" is among the cv terms, and *soccer* is in $document_i$ in $category_j$, then the algorithm counts the weight of *soccer*, and inserts the result in the matrix z . As a result, the feature matrix z is generated, with the columns containing feature terms, and the rows containing the documents. At the end of each row the category class label is included to provide the category annotation for the corresponding document. This annotated feature matrix is then used to derive the classification model by feeding it to any classifier of choice (e.g. SVM).

The following section demonstrates the application of the MFX algorithm to different datasets, and compares its performance to state of the art methods.

Algorithm 1 Proposed Feature Selection Algorithm-MFX

Input: c : Number of categories, d : Number of documents**Output:** z : Documents-Terms feature matrix

```
1: for  $cat = 1 \rightarrow c$  do
2:   for  $cat = 1 \rightarrow d$  do
3:     for  $j = doc + 1 \rightarrow d$  do
4:       Get all common terms between  $d_{doc}$  and  $d_j$ .
5:       Add the common terms to array  $cw_{cat}$ .
6:     end for
7:   end for
8:
9:    $\forall t_i \in cw_{cat}, z_{t_{ij}} = 10 * \frac{1 + \log(tf_{i,cw_{cat}})}{1 + \log(|cw_{cat}|)}$ 
10:  In matrix termSummary, update the cell that corresponds to term  $t_i$  and category  $cat$  to have the value  $z_{t_{ij}}$ . If  $t_i$  does not exist, add it and then update the cell value.
11: end for
12: Use the formulation in section 3.1 to find  $\delta$  and  $\tau_j$ , the optimum thresholds that maximize the distances between the categories (the columns in termSummary and minimize the distances between terms inside the same category
13: for  $\delta = 0.0 \rightarrow 1.0$  step: 0.1 arbitrary small step do
14:    $\forall t_i$  in termSummary
15:     1. Find  $max_i = max_w\_i - cat, cat = 1 \rightarrow c$ 
16:     2. Find  $median_i = median_w\_i - cat, cat = 1 \rightarrow c$ 
17:     3. if  $(\frac{max_i - median_i}{max_i} \leq \delta)$ , eliminate column  $t_i$  from termSummary
18:   for  $j = 1 \rightarrow c$  do
19:     for  $k = 1 \rightarrow c$  do
20:       Compute the objective function, and track the thresholds corresponding to the maximum of differences between intra-distances and inter-distances:
21:
22:       
$$max(\sum_{j=1}^c (\sum_{k=1, k \neq j}^c \frac{D_{jk}}{|c-1|} - d_j))$$

23:     end for
24:   end for
25: end for
```

Algorithm 2 Continue

19: output optimal δ and optimal $\tau_j \forall j$
 Store selected terms cw from all categories
20: **for** $cat = 1 \rightarrow c$ **do**
21: **for** $doc = 1 \rightarrow d$ **do**
22: $tf_{i,doc} \forall i \in cw$
23: Insert the calculated term in matrix z
24: **end for**
25: **end for**

Chapter 3

Results and Analysis

This chapter includes the experiments to evaluate the proposed feature selection method and its effect on text classification in general, and for emotion recognition in particular.

3.1 News Classification

In this section, we evaluate MFX on both Arabic and English documents. In every experiment, we (1) pick a dataset, (2) apply MFX, (3) feed the resultant *documentterm* training matrix to an SVM classifier, and (4) report the classification accuracy. We have used the central limit theorem to calculate the number of runs while achieving 90% confidence level with a precision value of 3%. The number of runs varied from 2 to 6 depending on the experiment. Therefore, we have performed 6 runs for all scenarios. The first two experiments target, in sections 3.1.1 and 3.1.2, Arabic documents. The third experiment, in section 3.1.3, targets large corpus of English documents, namely RCV1. The fourth and fifth experiments, in sections 3.1.4 and 3.1.5, target English documents on smaller sized corpora, and compare to a wide variety of conventional and modern feature selection methods.

3.1.1 MFX Evaluation on Small to Medium Sized Arabic Documents

The objective of the first experiment is to test the accuracy of MFX on Arabic documents when document sizes vary from small to medium and for different categories. Three types of datasets were used:

- *DS – CC* (Document Small - Category Close): Small number of documents with close (overlapping) categories

- $DS - CD$ (Document Small - Category Distinct): Small number of documents with distinct categories
- $DM - CD$ (Document Medium - Category Distinct): Medium number of documents with distinct categories

For DS-CC, the dataset was extracted from El Watan newspaper [42]. Four categories were selected: Culture, religion, sport, and economy, with each category having 50 documents and each document ranging in size from 2 Kilo Bytes (kB) to 3kB. Note that since culture often has religious elements, and religion often has impact on culture, we expected that these categories would have similar features leading to reduced classification accuracy in comparison to other categories.

For DS-CD, the dataset was also extracted from El Watan newspaper. Four categories were selected: Health, History, Sport, Economy, with each category having 50 documents ranging in sizes from 2kB to 3kB. We replaced the culture category with health and the religion category with history. The intuition behind this change was that replacing similar categories with more distinct ones should give higher classification accuracy as was confirmed in the experimental results.

For DS-CD, the dataset was extracted from CNN Arabic version [43]. Four categories were selected: Business, Sport, World, and Entertainment, with each category having different numbers of documents 50, 100, 150, and 200; and each document ranging in size from 3kB to 4kB. The objective behind this experiment is to test the classification accuracy when categories have different number of documents, and with different sizes.

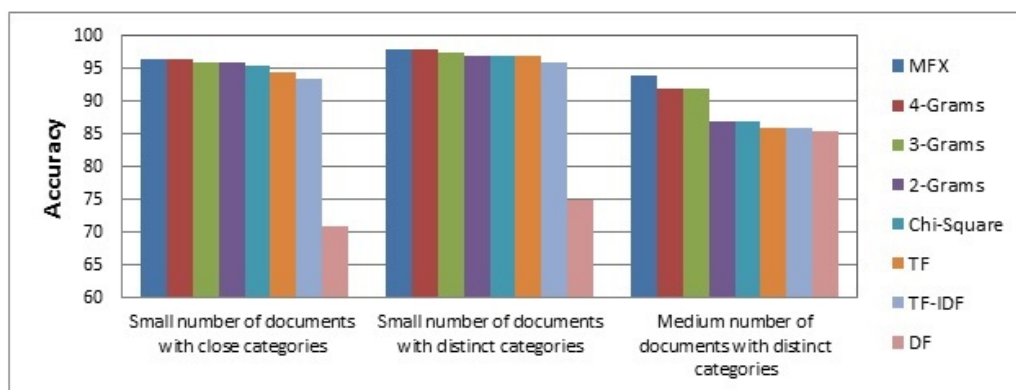


Figure 3.1: Classification accuracy when document sizes vary from low to medium and when categories are close and distinct

MFX, DF, TF, TF-IDF, N-character-Grams (2-Grams, 3-Grams, 4-Grams), and Chi-Squared statistics were then applied to the chosen datasets. We used these algorithms because of their effectiveness in feature selection [2]. SVM was used for document classification. The results of the classification accuracies are reported in Figure 3.1.

For DS-CC data (Document Small - Category Close), MFX achieved better accuracy than all the other methods, and equal accuracy to 4-grams. This can be explained by the fact that the majority of derived Arabic words were composed of 4 characters. Additionally, 4-grams acts like a light stemmer which means that the suffix of a given word is deleted, and if the same word is presented in normal format (4 characters), its frequency would increase. We also experimented with 5-grams and 6-grams, but it turned out that they just acted as term frequency (TF) since the words which have length greater than 5 were rare.

For DS-CD data (Document Small - Category Distinct), we noticed an increased accuracy for all the methods including MFX. The reason behind this increase in accuracy was that the content of the data affected the performance of the feature selection methods. Since the categories in this dataset were distinct, the classification accuracies improved for all methods. However, MFX continued to outperform the other methods under such scenarios.

For DM-CD (Document Medium - Category Distinct), the accuracy of the classification decreased for all tested feature selection algorithms, but MFX continued to produce the best classification accuracy. The accuracy drop in all methods is caused by the dataset contents and its unpredictability. Still, MFX outperformed the other approaches due to its discriminative features and the optimal choice of the number of features.

This whole experiment showed that MFX worked well for Arabic text under different dataset sizes and categories. It also showed that MFX was not affected by the number of documents. These results show that MFX addressed some of the challenges presented in [26] which stated that the data characteristics (number of documents, document size, and document content) can have an impact on the stability of accuracy results.

3.1.2 MFX Evaluation on Larger Sized Arabic Documents

The objective of this experiment is to show the accuracy of MFX on Arabic documents when the number of categories and the document sizes are relatively larger. Here we assume that 6kB sized documents are relatively large. This experiment compared MFX to SIPs which is the feature selection

method used by Amazon. The dataset was collected from Al-Jazeera [44] and OSCA [45]. Table 3.1 summarizes the dataset properties. Ten categories were used, with each category having different number of documents (50 to 100) and each document ranging in size from 2kB to 6kB.

Table 3.1: Dataset from Al- Jazeera newspaper and OSCA dataset

Category	Number of documents	Minimum size	Maximum size
Medical	50	2kB	6kB
Culture	55	2kB	6kB
Art	63	2kB	6kB
Law	66	2kB	6kB
Local	70	2kB	6kB
Society	75	2kB	6kB
International	85	2kB	6kB
Economy	92	2kB	6kB
Astronomy	94	2kB	6kB
Sport	100	2kB	6kB

The same feature selection methods used in the first experiment were also used for comparison in this experiment. Figure 3.2 shows the classification accuracy with the use of the classifier. In all the conducted experiments, MFX outperformed the other feature selection approaches.

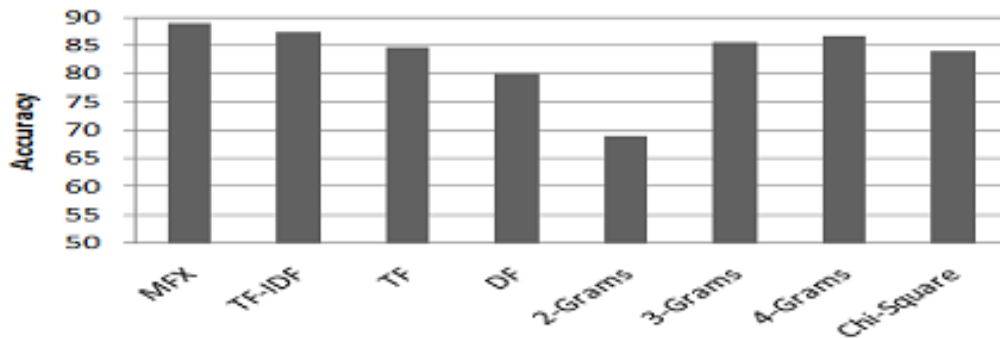


Figure 3.2: MFX comparison to other feature selection algorithms when applied to relatively large Arabic documents (6kB)

We also compared MFX to SIPs since it carried some resemblance with MFX. SIPs selects unique keywords for books to provide a summary of their content. The length of the phrases chosen for SIPs were 2 to 5 words, since the probability to have more than five words occurring together is rare. Figure 4 shows the accuracy results of MFX as being superior to SIPs.

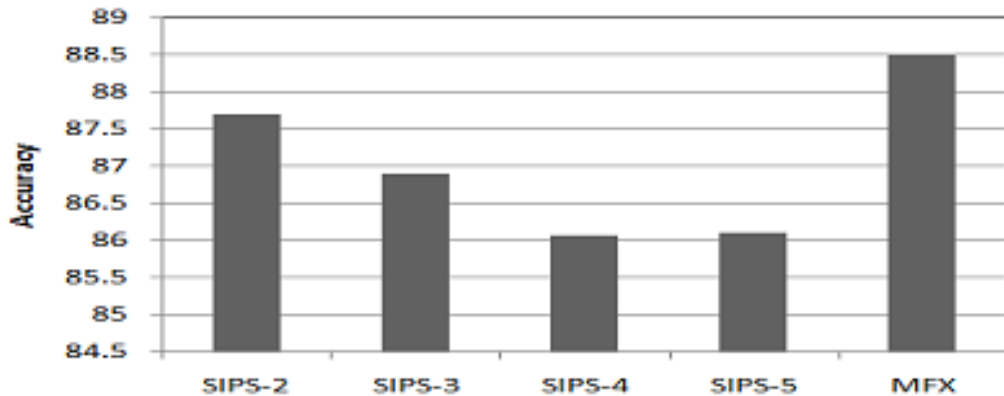


Figure 3.3: MFX comparison with SIPs when applied to relatively large Arabic documents

3.1.3 Benchmark evaluation with RCV1 dataset

The objective of this experiment is to test the performance of MFX on the full dataset provided by Reuters RCV1, which is considered a benchmark dataset for English text categorization [46]. In RCV1, each document belongs to one or more classes (categories). The classes are labeled based on a hierarchy tree from the more general classes to the most detailed ones. In our study, we used the most detailed classes which contain the largest number of categories (43 categories). The pre-processing step which mainly includes parsing the xml documents of RCV1 was done using Matlab. The purpose of this step is to obtain a clean dataset where each document belongs to one or more classes of the 43 categories; documents not belonging to any of the 43 categories are eliminated. Figure 3.4 shows the classification accuracy of MFX . TF and TF-IDF. TF and TF-IDF were selected for this experiment due to their recurrent use in text categorization. MFX outperformed TF and TF-IDF by 15% and 16% respectively. This shows the power of MFX when run on big dataset. Next, we evaluate MFX on a relatively smaller dataset, but considering more feature selection methods: conventional and modern.

3.1.4 MFX Evaluation on Reuters-21578 dataset with Conventional Methods

The objective of this experiment is to test the accuracy of MFX on English documents using the Reuters-21578 benchmark [47]. Four Categories were selected: ACQ, Corn, Crude, and Earn, with 331, 181, 330, and 330 documents respectively. The document sizes ranged from 1kB to 5kB. All the

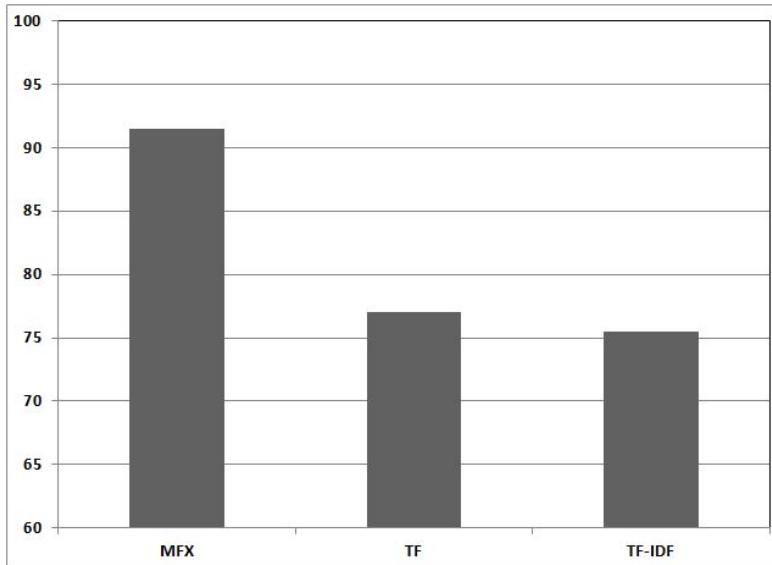


Figure 3.4: MFX comparison with conventional methods when applied to RCV1 dataset

feature selection methods used in the previous experiments were used in this experiment for comparison. Figure 3.5, and figure 3.6 show the resulting classification accuracies. Similar to when it was used with Arabic document, MFX outperformed the other methods when applied to English.

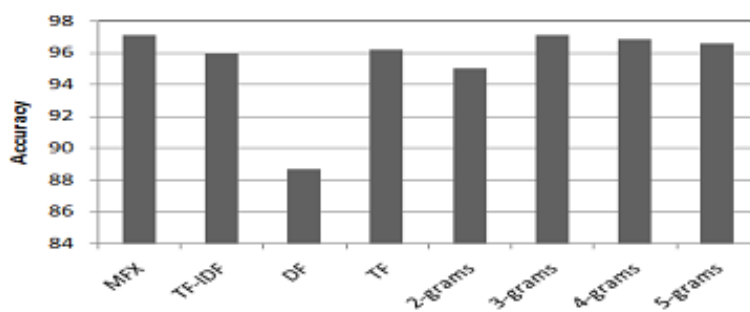


Figure 3.5: MFX comparison to conventional feature selection methods for English text

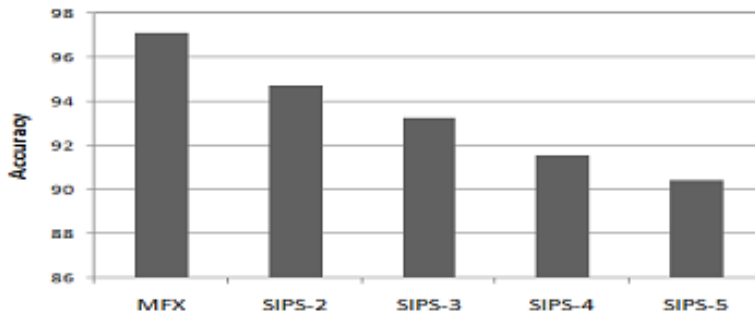


Figure 3.6: MFX comparison with SIPS when applied to English documents

3.1.5 MFX Evaluation on Reuters-21578 dataset with Modern Methods

The Purpose of this experiment is to compare MFX with modern feature selections methods such as Gini Index [17], OCFS [19], DIA [14], and CMFS [21]. We used the top 10 categories from Reuters. In order to provide fair comparison with the experimental results provided by [16], we only considered the top 2000 features selected by MFX (labeled as MFX-2000 in Figure 3.8) and compared with the other methodologies with respect to F1-micro measure (not precision). F1-micro measure is represented as follows:

$$P_{micro} = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} (TP_i + FP_i)} \quad (3.1)$$

$$R_{micro} = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} (TP_i + FN_i)} \quad (3.2)$$

$$F1_{micro} = \frac{2 * P_{micro} * R_{micro}}{R_{micro} + P_{micro}} \quad (3.3)$$

where TP, FP, and FN represent True Positives, False Positives, and False Negatives respectively. The results indicate an improvement of MFX in the range [6%-16%] over the other approaches.

3.2 Emotion Classification

In this section, we test the proposed method for emotion recognition from text. A brief description of the dataset is presented in 3.2.1 followed by the experiment results in section 3.2.2.

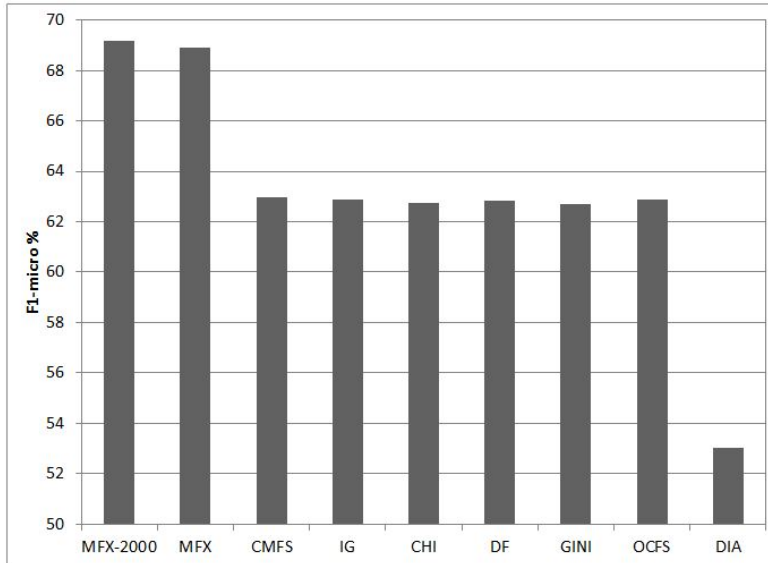


Figure 3.7: MFX comparison to modern classification methods for English text

3.2.1 Dataset used for Emotion Recognition from Text

ISEAR dataset: [48] consists of sentences, each annotated by the emotion conveyed by the speaker of the sentence. It captures situations that people have in their daily life which aroused the feeling of different emotions such as; *Joy, Fear, Anger, Sadness, Disgust, Shame, and guilt*. Table 3.2 provides a summary of the dataset. As we can the dataset is balanced.

Table 3.2: ISEAR summary dataset

Emotion	Number of sentences
Joy	1094
Fear	1095
Anger	1096
Sadness	1096
Disgust	1096
Shame	1096
Guilt	1093

Amman dataset: [49] it consists of 173 blog posts, consisting of 15205 sentences. Four judges were asked to label the dataset and also provide the intensity for the 7 emotional classes *Happiness*, *Sadness*, *Anger*, *Fear*, *Disgust*, *Surprise*, and *No Emotion*. Only the sentences which satisfied the complete agreement among the judges are selected. This results with a selection of 4090 sentences of the total (15205). Table 3.3 provides a summary of the Amman dataset. It can be seen that the data is unbalanced with almost 69% of the data belonging to the *NoEmotion* class.

Table 3.3: Amman summary dataset

Emotion	Number of sentences
Happiness	536
Fear	115
Anger	179
Sadness	173
Disgust	172
Surprise	115
No Emotion	2800

UIUC Children’s Story Corpus: [50] it consists of 176 children’s stories provided by three authors *Grim*, *Anderson*, and *Potter*. The annotation of this dataset was done by two annotators. The annotation process consisted of providing the emotion labels such as; *Happiness*, *Sadness*, *Fear*, *Anger*, *Surprise +*, *Surprise -*, *Disgust*, and *Neutral*. The polarities associated with the "Surprise" emotion were indicative of the mood: positive (+) or negative (-). The selection of the sentences was based on the complete agreement between the raters. Also, according to the data description file the following two emotional classes *Anger*, and *Disgust* are merged into one class *Anger* and similarly for *Surprise +*, and *Surprise -* classes are merged under *Surprise*. The final emotional classes were: *Happiness*, *Sadness*, *Fear*, *Anger*, *Surprise*, and *Neutral*. Table 3.4 provides a summary of the UIUC corpus.

3.2.2 Experiment results for Emotion Recognition

Several experiments were conducted using the dataset from the previous section. A comparison is also included with state of the art methods.

Table 3.4: UIUC Children’s Story Corpus

Emotion	Number of sentences
Happiness	445
Fear	166
Anger	218
Sadness	264
Surprise	114
Neutral	5621

3.2.2.1 Evaluation on Amman dataset

In this experiment, we have compared our method to the work done in [41]. After, applying MFX, we got the selectivity ratio δ equal to 0.12 and τ equal to 1. This meant that the best features set contained 4143 of the total number of features. Table 3.5, and table 3.6 present a list of the features based on the selection and the rejected features respectively.

Table 3.5: List of features selection based on δ

Term	Angry	Disgust	Fear	Happy	Surprise	Sad	No Emotion	δ
Angry	17	0	0	0	0	0	0	1
Awesome	0	0	0	19	0	0	0	1
Dislike	0	15	0	0	0	0	0	1
Lol	0	0	0	32	0	0	0	1
Scared	0	0	14	0	0	0	0	1
Love	0	0	1	50	0	3	1	0.96
Happy	1	0	0	26	0	2	1	0.942
Hate	1	15	1	0	0	0	0	0.9333
Good	0	1	1	42	0	3	11	0.928
Today	1	3	1	14	1	1	11	0.928
Comics	0	0	0	1	1	0	13	0.923
Guess	1	1	0	2	1	1	13	0.923
Work	1	4	1	5	2	2	25	0.92
Shit	11	1	1	0	1	0	0	0.77
Amazing	0	0	0	9	13	0	1	0.33

Table 3.6: List of rejected features based on δ

Term	Angry	Disgust	Fear	Happy	Surprise	Sad	No Emotion	δ
Accident	1	1	0	0	0	0	0	0
Acheive	1	1	0	0	0	0	0	0
Acoustic	1	1	0	0	0	0	0	0
Acquire	1	1	0	0	0	0	0	0

Table 3.5 presents the term occurrence of terms in the emotional classes. We notice that the terms selected indeed correspond to the emotion. For example, *Love* chosen for *Happy*, *Hate* chosen for *Disgust*, and *Shit* chosen for *Angry*. Table 3.6 shows a list of unwanted terms such as; *Accident*, and *Acoustic* which are not related to any of the emotional class. Moreover, the method showed some of the unique emotional terms such as: *Angry*, *Awesome*, and *Dislike*. Where they are presented only in one class. Figure 3.8 shows a comparison with the *Prior and Contextual Emotion* methodology [41]. Our proposed methodology shows that it can perform well even with a small set of data. We have selected the same subset as proposed in [41]. This subset consists of 1290 sentences out of 4090 sentences.

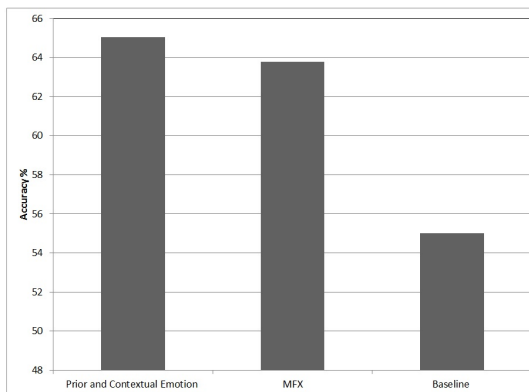


Figure 3.8: MFX Performance on Amman Dataset

Evaluation on UIUC Children’s Story Corpus dataset In this experiment, we have compared our method to the work done in [51]. The resulting selectivity ratio δ was equal to 0.03 and τ was equal to 1. The best features set contain 7240 features of the total number of features. Figure 3.9 shows

that MFX outperformed the method used in [51] their feature selection technique is based on scanning the documents based on *WordNet* only. While our method find other relation between term and the associated emotional class. such as; the term *King* belongs to *Angry*, *Happy*, *Sad*, *Surprise* respectively. Figure 3.10 presents the relation of word king with the emotional classes. For example, it is labeled 29 times as *Happy*.

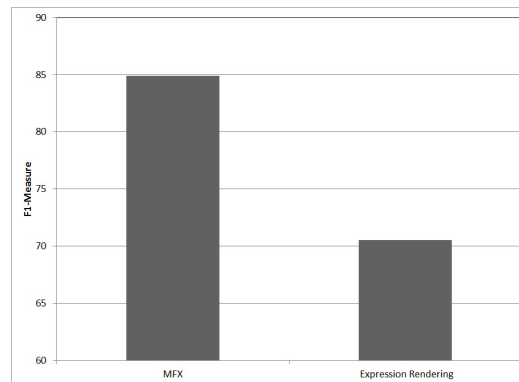


Figure 3.9: MFX Performance on UIUC Children’s Story Corpus Dataset



Figure 3.10: relation between term and the associated emotional class

3.2.2.2 Evaluation on ISEAR dataset

In this experiment we have compared our method to the work done in [37]. This paper compare several statistical approaches used in text mining such

as; *Unigram*, *Bigram*, and *Trigram*. In addition to methods that are based on *WordNet* lexicon. Also, we have compared to EmotiNet [52]. According to the experimental setup in paper [37] we have selected the same subset of ISEAR for comparison purposes. This subset consists of sentences that contain a descriptions of situations involving family members. The selectivity ratio δ was equal to 0.03 and the derived τ was equal to 1. The best features set contained 2390 features of the total number of features. Figure 4.1 shows the experimental result. MFX outperformed several methodologies such as: Unigram, WordNet plus linguistic inquiry. While the combination of Unigram and Bigram outperformed our method.

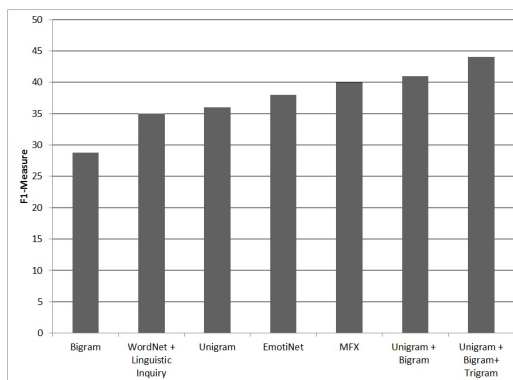


Figure 3.11: MFX Performance on ISEAR Dataset

3.3 Conclusion

In summary, MFX outperformed well-known and commonly used feature selection methods for English and Arabic text classification, under different sized corpora that include RCV1, Reuters-21578, and selections from Arabic newspapers that include CNN Arabic, Al-Jazeera, and El Watan. The method also showed robustness for different document sizes and close or distinct categories. The major advantage of MFX remains to be its consistency in outperforming the existing and mostly used feature selection approaches, allowing for MFX to be the feature selection method of choice regardless of the dataset or language. Concerning the conducted experiments on Emotion recognition from text, the method outperformed several methodologies while testing it on different Benchmark datasets such as: ISEAR, UIUC children stories, and Amman.

Chapter 4

Emotion Recognition from Physiological Data in Natural Settings

This Chapter present a method for emotion recognition from physiological data in natural settings. Section 4.2 presents the details of the proposed methodology for generating accurate ground truth physiological data, and its annotations with emotion labels. Section 4.3 describes the steps of the experiments conducted. A summary is presented at the end of the chapter highlighting the remaining open challenges as a future work.

4.1 Literature Review

We present in Table 4.1 a summary of most commonly used sensors and the corresponding bio-signals for collecting physiological measurements.

Table 4.1: Summary of Sensor and Corresponding Bio-Signals

Ref	Type	Sensors	Indicator of
[53]	Cardiovascular System	Heart rate (HR)	negatively valence emotion
[54]	Cardiovascular System	Electrocardiography (ECG) for Heart Rate Variability (HRV)	Stressed (indicated by Irregular HRV) Relaxed (indicated by stable HRV)
[55]	Cardiovascular System	Blood Pressure (BP)	Fear and Anger
[56]	Electro dermal Activity	Galvanic Skin Response (GSR)	Valence and Arousal of the Sympathetic Nervous System
[57]	Electro dermal Activity	Skin Conductance (SC)	Valence and Arousal of the Sympathetic Nervous System
[8]	Respiratory	Electrocardiography (ECG) for Respiratory Sinus Arrhythmia (RSA)	Valence and Arousal

It can be seen from Table 4.1 that Galvanic Skin Response *GSR* and Skin Conductance *SC* can be used to assess valence and arousal. The sensors can be installed on a persons fingers, and can be used to collect real time signals. These signals are used in our study of emotion recognition from Physiological data. Table 4.2 provides a survey of the methods used for emotion recognition.

Here is a description for the acronyms used in the table, within the different columns listed in table 4.2. For the sensors (second column) the acronyms used are as follow: *ECG*: electrocardiography, *EDA*: electrodermal activity, *EMG*: electromyogram, *EOG*: electrooculogram, *GSR*: Galvanic Skin Response, *HR*: Heart Rate, *RA*: respiration activity, *SC*: Skin Conductance, *BV*: Blood Volume, *T*: skin temperature, *P*:Pulse, *A*: Accelerometer. For the stimuli used(sixth column) the acronyms used *IAPS*= International Affective Picture System.

From Table 4.2, it can be seen that most of the methods relied on lab settings for stimulating emotions. While the accuracy may be high for these

Table 4.2: Survey of Methods for Emotion Recognition using Physiological Signal

<i>Ref</i>	<i>Sensors</i>	<i>Subject</i>	<i>Emotions</i>	<i>Annotation</i>	<i>Stimuli</i>	<i>Feature selection</i>
[7]	<i>EMG</i> <i>ECG</i> <i>SC</i> <i>RA</i>	3	Joy Anger Sad Pleasure	Music and memory recall	Predefined	Statistical Energy Sub and Spectrum Entropy
[58]	<i>ECG</i> <i>EDA</i> <i>T</i>	125	Sad Anger Stress Surprise	Predefined	Multimodal	Mean, Std-Dev Frequency Power
[10]	<i>RA</i> <i>SC</i> <i>T</i> <i>BV</i> <i>EMG</i>	1	Neutral Anger Hate Grief Platonic love Romantic love Joy Reverence		Personalized Imagery	Std-Dev absolute values of first and second derivative of raw signals
[35]	<i>GSR</i> <i>HR</i> <i>A</i>	19	Arousal Valence	Real life	Self- Assessment	Mean variance Median inter-quartile range
[59]	<i>EMG</i> <i>HR</i> <i>SC</i>	36	Arousal Valence	Robot	Self- Assessment	N/A
[60]	<i>ECG</i> <i>RA</i> <i>EDA</i>	9	Happiness Disgust Fear	IAPS	Predefined	Mean Std-Dev

methods, they do not reflect natural behavior. in natural settings.

The approaches used in emotion recognition consists of the following steps:

- a. Collecting/Generating Ground Truth Data for each modality
- b. Machine Learning Process : This represents the training phase of the learning process. Here, a mathematical model is built based the developed training data, often done in the lab. The model is then tested with parts of the training data.
- c. Evaluating Accuracy: the most common technique used for evaluation is the Leave- One -Out Cross-Validation (LOOCV). According to [61], there are two ways for classification: the dependent approach and the independent approach. In the dependent approach, the tuples corresponding to one participant are extracted for validation. In the independent approach, at each step of cross validation, the samples of one participant are taken out as test set while the remaining samples of other participants are used to train the classifier. The independent approach is the most commonly used for emotion recognition.

4.2 Proposed Approach for Emotion Recognition from Physiological Data in Natural Settings

For emotion recognition from physiological data, the objective is to develop a new accurate recognition model in natural settings. Our two main contributions are:

1. Generating accurate ground truth data with emotion label annotations for physiological data
2. A comparative study evaluation for most effective classification approach for emotion recognition from physiological data.

The methodology for generating accurate ground truth data is described in sub-section 4.2.1, and the methodology to conduct the study is presented in sub-section 4.2.2.

4.2.1 Method for Ground truth Collection in Natural Setting

The proposed methodology is an extension of a study that was started by [10]. An experiment was conducted with 13 participants for five days, where the participants were asked to annotate their emotion during the day on their mobile phones. The emotions were chosen on a three dimensional axes: Valence, Arousal, and Control. Simultaneously, the phone collected physiological data from the participants. The collected measurements included GSR and HR. The time stamps for the annotations and measurements were automatically logged by the phone application. The participants also had a choice to enter to their physical state and the type of activity they were doing (e.g. walking, deskwork, meeting,...,etc). At the end of each day, the participants were interviewed to get a recall of the main events during the whole day. A sample of the collected data is shown in Figure 4.1. The work in this thesis, extends the results of the conducted experiments by proposing a method to augment the collected annotation with more accurate labels. The idea is to have expert psychology evaluators review the evidence provided by the interviews, and the collected experiment data, and then apply their own assessment for the right annotations and time stamps. Three experts are proposed for this evaluation. Here are the steps followed by each of the experts:

1. Analyze the self-assessment annotation provided by the participants.
2. Analyze the interview reports for each day.
3. Provide a confidence level based on the evidence from the interviews for each of the emotion annotations provided by the participant.
4. Suggest a change in rating where confidence is low in participant self-annotation.

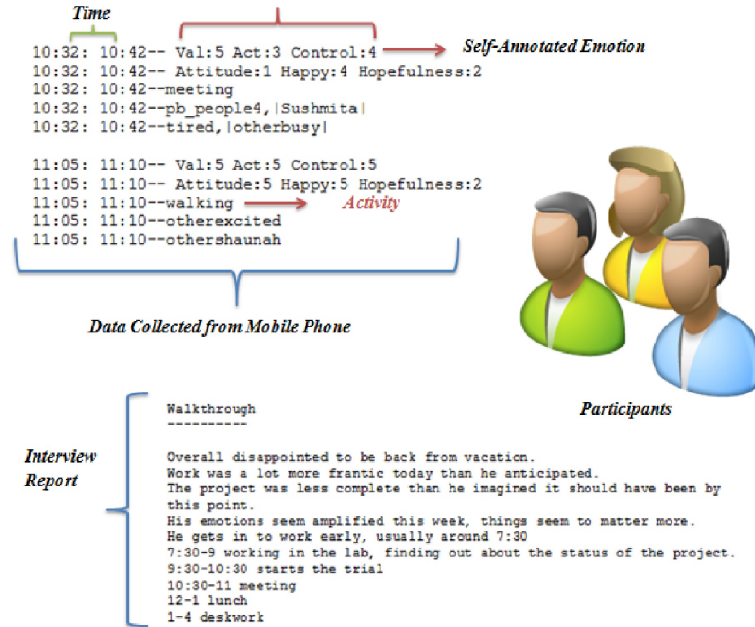


Figure 4.1: Data Collection Model

A template of the expected outcome is provided to each of the independent experts, and is shown in Figure 4.2. Once all the independent assessments are collected, the independent ratings are combined to yield a more accurate ground truth.

PARTICIPANT DATA AND EVALUATION				INDEPENDENT RATER ASSESSMENT AND EVALUATION																			
URL/ID	ACTIVITY	DEVELOPER	OS	NEW TIME				NEW VIEWS				NEW ACTIVATION				NEW DOMINANCE				NEW PHYSIOLOGICAL			
Condition	Activity	Developer	OS	Raw	Confid	Expans	Relax	Raw	Confid	Expans	Relax	Raw	Confid	Expans	Relax	Raw	Confid	Expans	Relax				
Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak	Peak			
Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable	Stable			
Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted	Adapted			
Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable	Unstable			

Figure 4.2: Proposed Template for Raters

4.2.2 Inter-Raters Agreement Calculation

There are different techniques used to test the inter-rater agreement such *Cohenskappa* [62], *Fleiss K* [63]. However, these have limitations. The *Cohenskappa* measure cannot handle more than two raters and *Fleiss K*

Table 4.3: Alpha Measures for Different Emotional Classes

Emotional Class	α
Valence	0.8722
Arousal	0.8576
Control	0.903

measure cannot handle nominal data. We have selected the Krippendorff α [64] method since it measures agreements for nominal, ordinal, interval, and ratio data, rendering the reliabilities for such data fully comparable across different metrics. In addition, it can calculate the agreement for more than two raters. The α measures were calculated by using SPSS software [65]. The results are summarized in table 4.3

The results showed that there is high correlation among raters for a good agreement the value of α should be greater than 0.5. According to table 4.3 all the measures were above this threshold. We have selected to choose the average as a measure to combine the raters assessments.

4.3 Evaluation and Ranking of Physiological Features for Emotion Recognition

To determine the most relevant physiological features for emotion recognition, we have conducted several feature selection experiments using forward selection, and backward elimination methods. The emotions were classified by using decision tree. Here are the steps followed:

- **Extract Features from Physiological Data:** The following features are derived from the raw physiological data at each relevant time stamp and corresponding window [66]:
 - **Running Mean:** The running mean computes a vector of mean values over time for a specific input signal by using a large rectangular window that is shifted across the feature vector. The size of the rectangular window depends on the input signal. This makes it possible to distinguish between phase, fast changes and tonic, and slow moving components in the analyzed signals. For example, Figure 4.3 shows a case where a participant participate in an event. At the beginning the participant was working by herself. Then, discussed issue with coworker beginning at the first circled

event. Then, she went to social event beginning at 2nd circled event. Later, they brought out cake at the 3rd circled event and participant annotated that she was very happy.

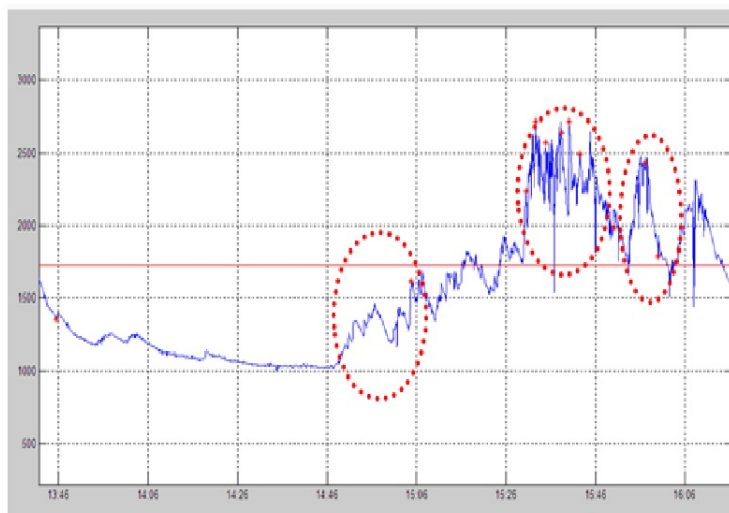


Figure 4.3: Participant Event

- **Running Standard deviation:** This procedure calculates the local standard deviation of the signal and results in a vector that could be described as changes in the current signal.
- **Slope:** The slope is an approximation of the first derivative and therefore indicates fast changes in the sensory data.
- **Combine** The generated features with the associated ground truth emotion labels.
- **Test** Apply feature reduction technique, then test with the decision tree classifier.
- **Analyze** Compare the effects of different features on classification accuracy.

4.4 Experiment for Creating Ground Truth Data

This section shows the results of the steps followed to create the ground truth data with labeled emotions, including: preprocessing, rater assessments, data cleaning, and choice of computed features.

Section 4.4.1 shows the results of class annotation based on the rater assessments. Section 4.4.2 presents the features extracted from physiological data followed by data mapping between the physiological features and the labeled emotional class in section 4.4.3. Finally, section 4.4.4 presents the emotional matrix model for *Valence*, *Arousal*, and *Control* respectively. Figure 4.4 presents the steps of the data processing phase.

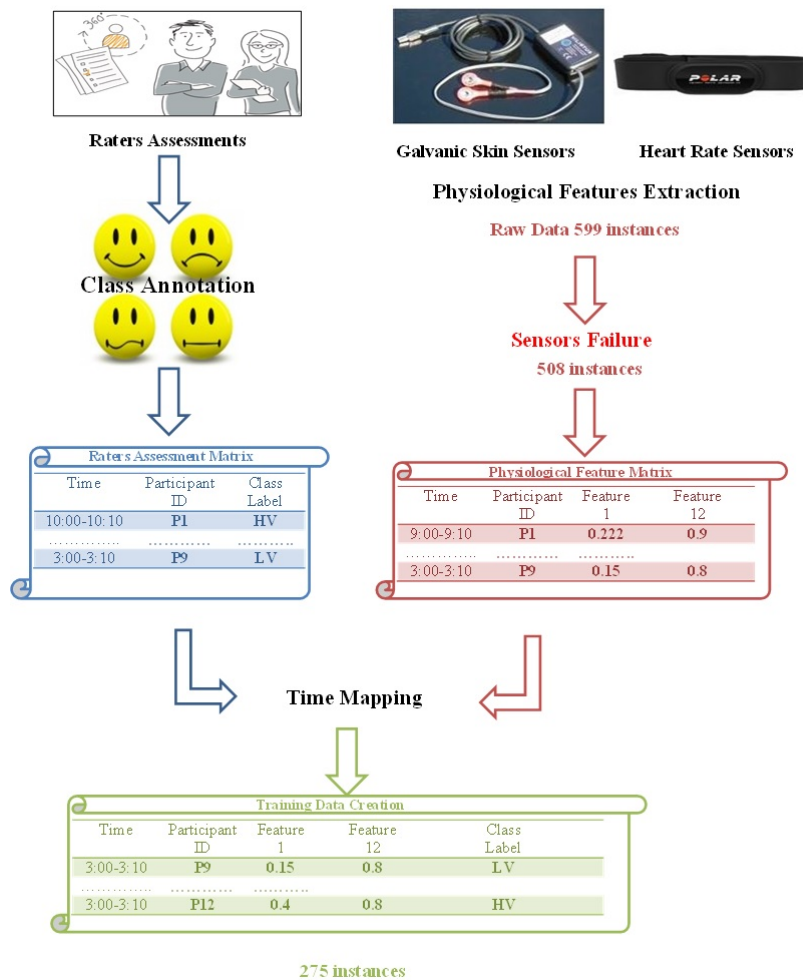


Figure 4.4: Preprocessing Phase

4.4.1 Data Labeling

The calculated average values of the raters assessments were calculated for each instance in each class of Valence, Arousal, and Control. Then, for each

class three groups (Low, Neutral, and High) were selected. For example, the emotional class *Valence* will be labeled as follow: *HV*, *NV*, and *LV*. Table 4.4 provides a summary of the classification of the instances. The

Table 4.4: Summary of the Classification of the Instances

Raters averaging value (X)	Class
$X < 3.8$	Low
$3.8 \leq X \leq 4.1$	Neutral
$X > 4.1$	High

rating of the emotion class is based on a seven-points Likert scale [67]. As shown in table 4.4, instances that receive a value between 3.8 and 4.1 are classified as neutral.

4.4.2 Physiological features extractions

In this study, we have used two physiological sensors which were GSR and HR. The features computed for these sensors are summarized in table 4.5.

Table 4.5: Features Extracted from the Physiological Sensors

Feature ID	Description of the feature	Sensor type
Feature 1	Means normalized with respect to the entire day	GSR
Feature 2		HR
Feature 3	Means normalized with respect to the beginning of the emotion period	GSR
Feature 4		HR
Feature 5	Variance normalized by dividing by the variance of the day	GSR
Feature 6		HR
Feature 7	Un-normalized variance	GSR
Feature 8		HR
Feature 9	The signal kurtosis	GSR
Feature 10		HR
Feature 11	The GSR peak detector of the max amplitude	GSR
Feature 12	The GSR peak detector of the max valley	GSR
Feature 13	The GSR peak detector of the peak frequency	GSR

4.4.3 Data Cleaning

Sometimes, for unknown reasons the sensors fails to store the physiological data. This resulted in a reduction of 15% from the original 599 instances. In order to map between the physiological features and the raters assessments, we used two factors which were the participant ID, and the starting time of the emotion. Also, The mapping process resulted in a reduction of 55% from the original 599 instances Figure 4.5 present a summary of the data reduction of the sensors failures and the time mapping process. Figure 4.6 shows how much data were left according to each participant.

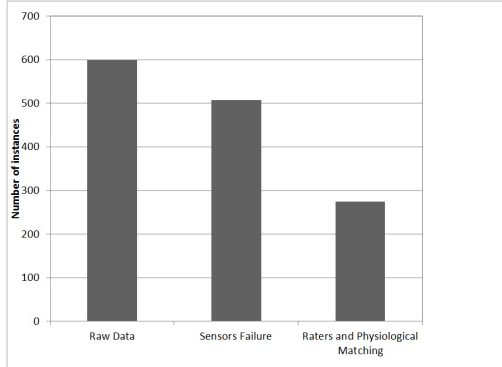


Figure 4.5: Data Reduction

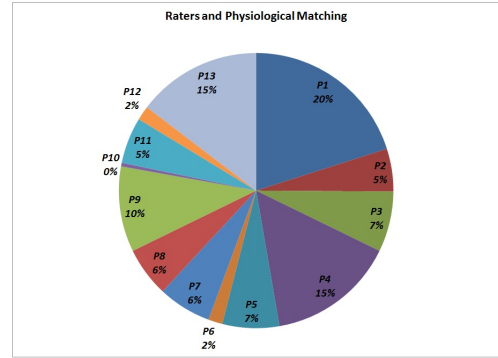


Figure 4.6: Participants Distribution

4.4.4 Training Data creation

Finally, after processing all the processes mentioned in subsection 4.4.1 to 4.4.3 we have created our training data for Valence, Arousal, and Control.

4.5 Observations

We have examined different feature reduction techniques such as: Backward Elimination, and Forward Elimination [68] to determine which features have the most impact on emotion recognition. Also, we have selected decision tree as a classifier for the whole experiments. We have selected three types of testing as follows:

1. **Raters assessment:** where each rater provides his or her own assessment and the average values among the raters was applied.
2. **Complete raters agreement:** is the case where only raters assessment with the complete matching rating values were chosen.
3. **Participant assessment:** where the participant provides his/her own rating.

4.5.1 Valence Evaluation

Figure 4.7 presents the results of applying the Valence training data to the scenarios listed above. The average raters assessments outperformed the participant assessment, and the complete agreement by 19% and 8% respectively. Table 4.6 presents a list of features selected by the forward and backward

elimination. Most of the features which correlate with Valence are related to the GSR sensors.

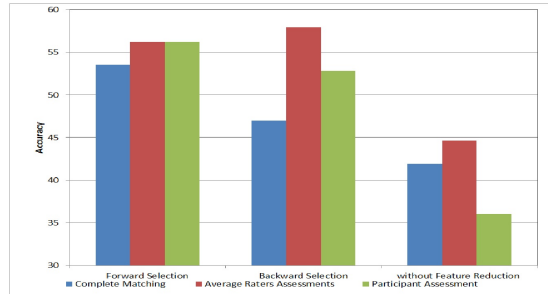


Figure 4.7: Evaluation of Valence Model

Table 4.6: Features Selection by Applying Forward and Backward Elimination for Valence

Feature	Forward	Backward	Sensor
Means normalized with respect to the entire day		X	GSR
Means normalized with respect to the beginning of the emotion period		X	GSR
Means normalized with respect to the beginning of the emotion period		X	HR
Variance normalized by dividing by the variance of the day		X	GSR
Un-normalized variance		X	GSR
The signal kurtosis		X	GSR
The GSR peak detector of the max valley	X	X	GSR
The GSR peak detector of the peak frequency		X	GSR

4.5.2 Arousal Evaluation

Figure 4.8 shows that the performance is almost the same for complete matching method and the average raters assessment. Similar to the previous experiment, the participant assessment had the lowest accuracy compared to

others. In addition, the forward selection methodology outperformed the backward selection by 13% even with a lower number of features as shown in table 4.7.

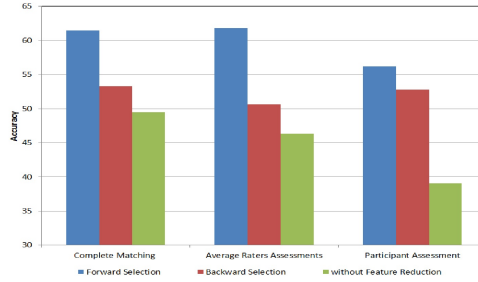


Figure 4.8: Evaluation of Arousal Model

Table 4.7: Features Selection by Applying Forward and Backward Elimination for Arousal

Feature	Forward	Backward	Sensor
Means normalized with respect to the entire day		X	GSR
Means normalized with respect to the entire day		X	HR
Means normalized with respect to the beginning of the emotion period		X	GSR
Means normalized with respect to the beginning of the emotion period		X	HR
Variance normalized by dividing by the variance of the day		X	HR
Un-normalized variance		X	GSR
The signal kurtosis		X	GSR
The signal kurtosis	X	X	HR
The GSR peak detector of the max amplitude		X	GSR
The GSR peak detector of the max valley		X	GSR
The GSR peak detector of the peak frequency		X	GSR

4.5.3 Control Evaluation

Similar to the previous experiments the forward elimination outperformed the backward technique by 18% as shown in figure 4.9. It seems that both Control and Arousal are correlated with the *kurtosis signal* of the Heart rate sensors. Table 4.8 presents a list of features selected by the forward and backward elimination.

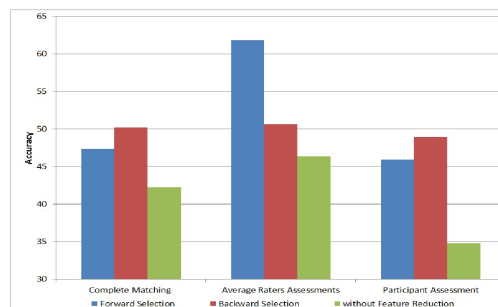


Figure 4.9: Evaluation of Control Model

Table 4.8: Features Selection by Applying Forward and Backward Elimination for Control

Feature	Forward	Backward	Sensor
Means normalized with respect to the entire day		X	GSR
Means normalized with respect to the entire day		X	HR
Means normalized with respect to the beginning of the emotion period		X	GSR
Means normalized with respect to the beginning of the emotion period		X	HR
Variance normalized by dividing by the variance of the day		X	HR
Un-normalized variance		X	GSR
The signal kurtosis		X	GSR
The signal kurtosis	X	X	HR
The GSR peak detector of the max amplitude		X	GSR
The GSR peak detector of the max valley		X	GSR
The GSR peak detector of the peak frequency		X	GSR

4.6 Conclusion

In the experiments we have examined different scenarios which are listed as follow: Raters Assessment, Participant Assessment, and Complete Raters Agreement. The results showed that there is a significant advantage in applying the Raters Assessment instead of Participant Assessment. Moreover, by applying feature reduction methodologies such as Forward and backward elimination we discovered the features that correlate with each emotions. For example, the features extracted from the GSR sensor correlate more with Valence.

Chapter 5

Conclusion

In this work , we have proposed a new algorithm for emotion recognition from text. The method is formulated as an to solve the optimization problem for feature selection from text. The method was tested in comparison with general-purpose text classification approaches using benchmark data (MFX) outperformed the state of the art methods for English and Arabic datasets such as: *RCV1* and *Reuters – 21758*. The results were also promising when applied to emotion textual datasets. Concerning emotion recognition from physiological data we addressed the problem related to generating accurate ground truth data in natural settings. Three psychology students analyzed the data provided by the participants and applied their own assessments. Experimental results showed that raters assessments outperformed the self annotations done by participants.

This work can be extended in many directions:

- Emotion Recognition from text
 1. In addition to the statistical measures of feature selection like MFX. One approach is to add additional features extraction based on sentence level such as: PoS tag, relations.
 2. Evaluate other discriminative functions : MI, Conditional Probabilities for the selectivity ratio.
- Emotion Recognition from Physiological Data.
 1. Implementation of systems for real life use.
 2. Development of universal Emotion recognition model not costume to specific users.
 3. Development of evolutionary Emotion Recognition model.

- Fusion of Emotion Recognition from text and Physiological data.
 1. Provide an experiment that can handle both textual and physiological data.
 2. Test MFX for extracting contextual features from the fusion model.

Bibliography

- [1] G. Ball and J. Breese, “Emotion and personality in a conversational agent,” *Embodied conversational agents*, pp. 189–219, 2000.
- [2] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin, “Emotion recognition from text using semantic labels and separable mixture models,” *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, no. 2, pp. 165–183, 2006.
- [3] E.-C. Kao, C.-C. Liu, T.-H. Yang, C.-T. Hsieh, and V.-W. Soo, “Towards text-based emotion detection a survey and possible improvements,” in *Information Management and Engineering, 2009. ICIME’09. International Conference on*, pp. 70–74, IEEE, 2009.
- [4] Z. Teng, F. Ren, and S. Kuroiwa, “Retracted: recognition of emotion with svms,” in *Computational Intelligence*, pp. 701–710, Springer, 2006.
- [5] Y. Y. Mathieu, “Annotation of emotions and feelings in texts,” in *Affective Computing and Intelligent Interaction*, pp. 350–357, Springer, 2005.
- [6] A. Balahur, J. M. Hermida, and A. Montoyo, “Detecting implicit expressions of emotion in text: A comparative analysis,” *Decision Support Systems*, vol. 53, no. 4, pp. 742–753, 2012.
- [7] K. H. Kim, S. Bang, and S. Kim, “Emotion recognition system using short-term monitoring of physiological signals,” *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.
- [8] J. Kim and E. André, “Emotion recognition based on physiological changes in music listening,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 12, pp. 2067–2083, 2008.
- [9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis; using physiological signals,” *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 18–31, 2012.

- [10] J. Healey, L. Nachman, S. Subramanian, J. Shahabdeen, and M. Morris, "Out of the lab and into the fray: Towards modeling emotion in everyday life," in *Pervasive computing*, pp. 156–173, Springer, 2010.
- [11] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, vol. 97, pp. 412–420, 1997.
- [12] D. Fragoudis, D. Meretakakis, and S. Likothanassis, "Best terms: an efficient feature-selection algorithm for text categorization," *Knowledge and Information Systems*, vol. 8, no. 1, pp. 16–33, 2005.
- [13] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.
- [14] S. S. Mengle and N. Goharian, "Ambiguity measure feature-selection algorithm," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 1037–1050, 2009.
- [15] D. Koller and M. Sahami, "Toward optimal feature selection," 1996.
- [16] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of machine learning research*, vol. 3, pp. 1289–1305, 2003.
- [17] H. Park, S. Kwon, and H.-C. Kwon, "Complete gini-index text (git) feature-selection algorithm for text classification," in *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*, pp. 366–371, IEEE, 2010.
- [18] H. Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm," *Knowledge-Based Systems*, vol. 24, no. 7, pp. 1024–1032, 2011.
- [19] J. Yan, N. Liu, B. Zhang, S. Yan, Z. Chen, Q. Cheng, W. Fan, and W.-Y. Ma, "Ocfs: optimal orthogonal centroid feature selection for text categorization," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 122–129, ACM, 2005.
- [20] S. Qu, S. Wang, and Y. Zou, "Improvement of text feature selection method based on tfidf," in *Future Information Technology and Management Engineering, 2008. FITME'08. International Seminar on*, pp. 79–81, IEEE, 2008.

- [21] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing & Management*, vol. 48, no. 4, pp. 741–754, 2012.
- [22] Amazon, "§sipsĩ [online] available: <http://stackoverflow.com/questions/2009498/how-does-amazons-statistically-improbable-phrases-work>,"
- [23] H. Al-Mubaid and S. A. Umair, "A new text categorization technique using distributional clustering and learning logic," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 9, pp. 1156–1165, 2006.
- [24] F. Harrag, E. El-Qawasmeh, and P. Pichappan, "Improving arabic text categorization using decision trees," in *Networked Digital Technologies, 2009. NDT'09. First International Conference on*, pp. 110–115, IEEE, 2009.
- [25] H. M. Noaman, S. Elmougy, A. Ghoneim, and T. Hamza, "Naive bayes classifier based arabic document categorization," in *Informatics and Systems (INFOS), 2010 The 7th International Conference on*, pp. 1–5, IEEE, 2010.
- [26] R. Duwairi, M. Al-Refai, and N. Khasawneh, "Stemming versus light stemming as feature selection techniques for arabic text categorization," in *Innovations in Information Technology, 2007. IIT'07. 4th International Conference on*, pp. 446–450, IEEE, 2007.
- [27] R. Al-Shalabi and M. Evens, "A computational morphology system for arabic," in *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pp. 66–72, Association for Computational Linguistics, 1998.
- [28] A. Mesleh and G. Kanaan, "Support vector machine text classification system: Using ant colony optimization based feature subset selection," in *Computer Engineering & Systems, 2008. ICCES 2008. International Conference on*, pp. 143–148, IEEE, 2008.
- [29] A. M. A. Mesleh, "Chi square feature extraction based svms arabic language text categorization system.," *Journal of Computer Science*, vol. 3, no. 6, 2007.
- [30] W. B. Cavnar, J. M. Trenkle, *et al.*, "N-gram-based text categorization," *Ann Arbor MI*, vol. 48113, no. 2, pp. 161–175, 1994.

- [31] H. Binali and V. Potdar, “Emotion detection state of the art,” in *Proceedings of the CUBE International Information Technology Conference*, pp. 501–507, ACM, 2012.
- [32] F. Keshtkar and D. Inkpen, “A corpus-based method for extracting paraphrases of emotion terms,” in *Proceedings of the NAACL HLT 2010 Workshop on Computational approaches to Analysis and Generation of emotion in Text*, pp. 35–44, Association for Computational Linguistics, 2010.
- [33] E. Riloff, J. Wiebe, and T. Wilson, “Learning subjective nouns using extraction pattern bootstrapping,” in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pp. 25–32, Association for Computational Linguistics, 2003.
- [34] R. Burget, J. Karasek, and Z. SMÉKAL, “Recognition of emotions in czech newspaper headlines,” *Radioengineering*, vol. 20, no. 1, pp. 39–47, 2011.
- [35] J. C. de Albornoz, L. Plaza, and P. Gervás, “Improving emotional intensity classification using word sense disambiguation,” *Research in Computing Science*, vol. 46, pp. 131–142, 2010.
- [36] C. O. Alm, D. Roth, and R. Sproat, “Emotions from text: machine learning for text-based emotion prediction,” in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 579–586, Association for Computational Linguistics, 2005.
- [37] A. Balahur, J. M. Hermida, and A. Montoyo, “Detecting implicit expressions of emotion in text: A comparative analysis,” *Decision Support Systems*, vol. 53, no. 4, pp. 742–753, 2012.
- [38] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp, “Emotion recognition in text for 3-d facial expression rendering,” *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 544–551, 2010.
- [39] C. Strapparava and A. Valitutti, “Wordnet affect: an affective extension of wordnet.,” in *LREC*, vol. 4, pp. 1083–1086, 2004.
- [40] A. Balahur, J. M. Hermida, A. Montoyo, and R. Muñoz, “Emotinet: a knowledge base for emotion detection in text built on the appraisal theories,” in *Natural Language Processing and Information Systems*, pp. 27–39, Springer, 2011.

- [41] D. Ghazi, D. Inkpen, and S. Szpakowicz, "Prior and contextual emotion of words in sentential context," *Computer Speech & Language*, vol. 28, no. 1, pp. 76 – 92, 2014.
- [42] Watandataset, "[online] available: <http://www.watan.com>,"
- [43] CNN, "[online] available: <http://www.arabic.cnn.com>,"
- [44] ALjazeera, "[online] available: <http://www.aljazeera.com>,"
- [45] M. K. Saad and W. Ashour, "Osac: Open source arabic corpora," in *6th ArchEng Int. Symposiums, EEECS*, vol. 10, 2010.
- [46] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *The Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.
- [47] D. D. lewis, "Ťreuters datasetŤ, [online] available: <http://www.daviddlewis.com/resources/testcollections/reuters21578/>,"
- [48] K. Scherer and H. Wallbott, "The isear questionnaire and codebook," *Geneva Emotion Research Group*, 1997.
- [49] S. S.Aman, "Identifying expressions of emotion in text," in *Text, Speech and Dialogue* (Springer-Verlag, ed.), p. 196 Ū 205., 2007.
- [50] C. O. Alm, "Affect in text and speech," *Ph.D. dissertation, Univ. Illinois at Urbana-Champaign, Urbana*, 2008.
- [51] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp, "Emotion recognition in text for 3-d facial expression rendering," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 544–551, 2010.
- [52] A. Balahur, J. M. Hermida, and A. Montoyo, "Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 88–101, 2012.
- [53] R. R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions*. Prentice-Hall, Inc, 1996.
- [54] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability standards of measurement, physiological interpretation, and clinical use," *European heart journal*, vol. 17, no. 3, pp. 354–381, 1996.

- [55] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.
- [56] C. Zong and M. Chetouani, "Hilbert-huang transform based physiological signals analysis for emotion recognition," in *Signal Processing and Information Technology (ISSPIT), 2009 IEEE International Symposium on*, pp. 334–339, IEEE, 2009.
- [57] E. L. van den Broek, V. Lisỳ, J. H. Janssen, J. H. Westerink, M. H. Schut, and K. Tuinenbreijer, "Affective man-machine interface: unveiling human emotions through biosignals," in *Biomedical Engineering Systems and Technologies*, pp. 21–47, Springer, 2010.
- [58] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [59] G. Rigas, C. D. Katsis, G. Ganiatsas, and D. I. Fotiadis, "A user independent, biosignal based, emotion recognition method," in *User Modeling 2007*, pp. 314–318, Springer, 2007.
- [60] C. A. Frantzidis, C. Bratsas, M. A. Klados, E. Konstantinidis, C. D. Lithari, A. B. Vivas, C. L. Papadelis, E. Kaldoudi, C. Pappas, and P. D. Bamidis, "On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 14, no. 2, pp. 309–318, 2010.
- [61] G. C. Cawley, "Leave-one-out cross-validation based model selection criteria for weighted ls-svms," in *Neural Networks, 2006. IJCNN'06. International Joint Conference on*, pp. 1661–1668, IEEE, 2006.
- [62] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.," *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.
- [63] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.
- [64] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.

- [65] S. J. Coakes and L. Steed, *SPSS: Analysis without anguish using SPSS version 14.0 for Windows*. John Wiley & Sons, Inc., 2009.
- [66] A. Haag, S. Goronzy, P. Schaich, and J. Williams, “Emotion recognition using bio-sensors: First steps towards an automatic system,” in *Affective dialogue systems*, pp. 36–48, Springer, 2004.
- [67] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.
- [68] S. Derksen and H. Keselman, “Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables,” *British Journal of Mathematical and Statistical Psychology*, vol. 45, no. 2, pp. 265–282, 1992.