



MULTIMODAL ARABIC SENTIMENT ANALYSIS FOR VIDEO OPINION MINING

BY

SADAM HUSSEIN MOHAMMED AL-AZANI

A Dissertation Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

In

COMPUTER SCIENCE

MAY 2019

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This dissertation, written by **SADAM HUSSEIN MOHAMMED AL-AZANI** under the direction of his dissertation adviser and approved by his dissertation committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE**.

Dissertation Committee

Sayed 29 7/2019
Prof. El-Sayed M. El-Alfy (Adviser)

[Signature]
Prof. Shokri Selim (Member)

[Signature]
Prof. Amir Hussain (Member)

Tarek Helwa
Dr. Tarek El-Bassuny (Member)

[Signature]
Dr. Mohammad Alshayeb (Member)

[Signature]
Dr. Hamoud Aljamaan
Department Chairman

[Signature]
Prof. Salam A. Zummo
Dean of Graduate Studies

31/7/2019
Date



© Saddam Hussein Mohammed Al-Azani
2019

Dedication

With all of my love to my father, my mother, my wife and my children.

ACKNOWLEDGMENTS

In the name of Allah, Most Kind, Most Merciful. All Praise and thanks be to Allah alone, and His peace and blessings be upon His messenger and bondman our Prophet Muhammad, his family, his companions and his followers.

I wish to express my appreciation to Prof. El-Sayed M. El-Alfy, who served as my major advisor, for his guidance through the dissertation; his continuous support and encouragement can never be forgotten. My gratitude is also due to the dissertation committee members, Prof. Shokri Selim, Prof. Amir Haussain, Dr. Tarek El-Bassuny, and Dr. Mohammad Alshayeb, for their attentions and enlightening comments.

I would like to thank King Fahd University of Petroleum & Minerals (KFUPM) for supporting this research. I also would like to thank Tamar University, which gave me the opportunity for completing my PhD degree.

I would like to thank my parents whose prayers are always a great source of strength for me. I also want to thank my dear wife and children: Anwaar, Alhussein, Abrar and Amged, for all their patience and taking all the pain to give me time to finish this work. Last, but not least, thanks for everyone who helped me.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	vi
LIST OF TABLES	xi
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xxiii
ABSTRACT (ENGLISH)	xxi
ABSTRACT (ARABIC)	xxiii
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement and Motivation	3
1.2 Research Objectives	6
1.3 Scope of Work	7
1.4 Research Design	8
1.5 Dissertation Contributions	11
1.6 Dissertation Organization	17
CHAPTER 2 BACKGROUND AND LITERATURE REVIEW	19
2.1 Text-based Sentiment Analysis	20
2.1.1 Text-based Sentiment Analysis Approaches	20
2.1.2 Imbalance Class Problem in SA	24
2.1.3 Deep Learning based Approaches for SA	27
2.2 Opinion Spam Detection Approaches	30

2.2.1	Non-Arabic Opinion Spam Detection	32
2.2.2	Arabic Opinion Spam Detection	34
2.3	Multimodal Sentiment Analysis	37
2.3.1	Fusion Approaches	38
2.3.2	A-T Sentiment Analysis	41
2.3.3	T-V Sentiment Analysis	42
2.3.4	A-V Sentiment Analysis	47
2.3.5	A-T-V Sentiment Analysis	49
2.3.6	Multimodal Sentiment Analysis Datasets	53
2.4	Discussions	58
2.5	Summary	64
CHAPTER 3 TEXTUAL ARABIC SENTIMENT ANALYSIS		68
3.1	Framework of Textual Arabic SA	69
3.2	Feature Extraction	69
3.2.1	Term Frequency-Inverse Document Frequency (<i>tf-idf</i>)	70
3.2.2	Latent Semantic Analysis (LSA)	71
3.2.3	Word Embeddings	72
3.3	Classification Methods	75
3.3.1	Single Classifiers	75
3.3.2	Ensemble Learning	82
3.3.3	Deep Learning	84
3.4	Handling Class Imbalance Problem	88
3.5	Performance Evaluation	91
3.5.1	Evaluation methods	91
3.5.2	Evaluation metrics	93
3.6	Experiments	96
3.6.1	Experimental Settings	96
3.6.2	Feature Techniques Evaluation	99
3.6.3	Oversampling Techniques Evaluation	100

3.6.4	Deep Learning Techniques Evaluation	102
3.7	Summary	104
CHAPTER 4 TEXTUAL-EMOJIS SENTIMENT ANALYSIS		106
4.1	Background	107
4.1.1	Emoticons vs. Emojis	107
4.1.2	Emojis in Social Media	108
4.1.3	Emojis in Sentiment Analysis	114
4.2	Emojis-based Sentiment Classification	114
4.2.1	Dataset Preparation	115
4.2.2	Emojis-based Features	117
4.2.3	Fusion with Textual Features	122
4.3	Experiments	126
4.3.1	Results and Discussion	126
4.3.2	Handling Emojis Class Imbalance Issue	130
4.4	Summary	134
CHAPTER 5 MULTIMODAL SENTIMENT ANALYSIS		139
5.1	Multimodal Sentiment Analysis Framework	139
5.2	Dataset Preparation and Collection	143
5.3	Feature Extraction	146
5.3.1	Acoustic Features	146
5.3.2	Transcribed Textual Features	147
5.3.3	Visual Features	147
5.4	Multimodal Sentiment Analysis	152
5.4.1	Single-level Fusion	152
5.4.2	Multi-level Hybrid Fusion	154
5.5	Experiments and Results	155
5.5.1	Experimental Settings	155
5.5.2	Unimodal Results	156
5.5.3	Single-level Fusion Results	157

5.5.4	Multi-level Hybrid Fusion Results	157
5.5.5	Enhancement of Visual Features	158
5.5.6	Analysis and Discussion	162
5.6	Summary	163
CHAPTER 6 EFFECTS OF DEMOGRAPHICS ON SENTIMENT		
	ANALYSIS	165
6.1	Demographic Detection	166
6.2	Multi-class Demographic with SA	169
6.3	Multi-label Demographic and SA	171
6.4	Effects of Demographics	176
6.5	Enhanced Multi-class Demographic with SA	180
6.6	Summary	181
CHAPTER 7 CONCLUSIONS AND FUTURE DIRECTIONS		183
7.1	Contributions	186
7.2	Challenges and Limitations	189
7.3	Future Directions	190
APPENDIX A LIST OF PUBLICATIONS		191
A.1	Conferences	191
A.2	Journals	193
A.3	Patents	193
APPENDIX B STATISTICAL TESTS		194
REFERENCES		198
VITAE		234

LIST OF TABLES

2.1	Opinion spam detection approaches	36
2.2	Multimodality and tasks	38
2.3	Fusion Levels	40
2.4	Multimodal Datasets	59
2.5	Multimodal sentiment analysis level	60
2.6	Algorithms	65
2.7	Results and Affects “↑” means positive effect, “↓” means negative effect and “=” means no effect	66
3.1	Confusion matrix definition	93
3.2	Description of the datasets used in the experimental study	96
3.3	Summary of evaluated base and ensemble classifiers	98
3.4	Training parameters of Arabic word embeddings	98
3.5	Performance of <i>tf-idf</i> and LSA features for ASTD dataset.	99
3.6	Word embedding based features performance using ASTD dataset. The highest results are shown in bold font	100
3.7	Traditional features performance using GS-dataset	100
3.8	Word embedding based features performance using GS-dataset	101
3.9	The effects of over-sampling techniques for ASTD	101
3.10	The effects of over-sampling techniques for Gold-Standard dataset	101
3.11	The effects of over-sampling techniques for higher imbalance ratio ASTD dataset	102

3.12	Performance comparison of various models on ASTD and ArTwitter datasets with static and non-static initializations for CBOW and skip-gram word embeddings	103
3.13	Compilation optimizers with ArTwitter and non-static CBOW model .	103
3.14	Comparisons with other related approaches	104
4.1	Description of the evaluated dataset showing the various sources and number of instances that contain emojis	116
4.2	Training parameters emojis embeddings	121
4.3	Performance comparison of ten machine-learning approaches using textual features extracted by five different methods; the highest results are marked in bold.	127
4.4	Results using emojis based features	129
4.5	Fusion of Word2Vec CBOW, Word2Vec Skip Gram and Emojis frequencies at feature, score and decision levels	130
4.6	Summary of the used datasets with different imbalance ratio	133
4.7	Results using the original dataset (Dataset I)	134
4.8	Results using highly imbalanced dataset (Dataset II)	135
4.9	Results using the more highly imbalanced dataset (Dataset III)	136
4.10	Results using the more highly imbalanced dataset (Dataset IV)	137
5.1	SADAM dataset statistics	142
5.2	SADAM dataset description	142
5.3	Description of the considered speakers' age-groups	145
5.4	Consumed time for visual feature extraction for all videos	151
5.5	The size of feature vectors	154
5.6	Unimodal systems (Baseline).	156
5.7	Single-level fusion of feature, score and decision techniques	158
5.8	Multi-level hybrid fusion of feature, score, and decision fusion techniques	159
5.9	Multimodal sentiment analysis with the hybrid visual features	160
5.10	Benchmarking our results	163

6.1	Results for demographic recognition systems	167
6.2	Results for multi-class demographic with sentiments recognition systems	170
6.3	Results demographic as a new modality with its effects on other modalities	180
6.4	Results for multimodal gender, age and sentiment recognition systems when applying HOF and HOG	181

LIST OF FIGURES

1.1	A taxonomy of sentiment analysis	3
1.2	The high-level architecture of the methodology.	11
2.1	Type of machine supervised learning approaches	28
2.2	Multimodal sentiment analysis taxonomy	38
2.3	Fusion Levels	39
2.4	A tri-modal sentiment analysis system showing the fusion at: (a) fea- ture level, (b) score level and (c) decision level	39
3.1	Layout of the textual Arabic SA approach	69
3.2	Neural network architectures for learning word2vec models (left: CBOW, right: Skip-gram	74
3.3	An example of sentence representation using word embeddings	75
3.4	Hyperplanes and margins in SVM	78
3.5	Adopted CNN architecture for Arabic sentiment analysis	86
3.6	Combined LSTMs for Arabic sentiment analysis	88
3.7	(a): Distribution of imbalanced dataset (b) Synthetic examples gener- ated using SMOTE	90
3.8	Example of an Arabic stop-word with its different forms through pre- fixes, suffixes and affixes	97
3.9	The evaluated deep learning models per each dataset.	99
4.1	Emojis taxonomy	108
4.2	Example of “cow face ” emoji appearance on different platforms	111

4.3	A general overview of the proposed approach (a) the basic model of single modalities of text and emojis (b) the fusion model of both modalities at different levels	115
4.4	The top-ranked 30 emojis in the dataset (sorted by the score computed by (a) ReliefF algorithm, (b) by CAE)	118
4.5	Top 10 emojis in ESR	120
4.6	Different queries for emoji2vec using CBOW and skip-grams	122
4.7	An example of sentence representation using emoji2vec embeddings	123
4.8	Performance comparison using ROC curves and AUCs for five sentiment classification models using SVM	131
4.9	Layout of the proposed approach for handling emojis imbalance issue	132
4.10	ROC and AUC for Dataset I for the single classifier, conventional bagging classifier and balanced bagging classifier	135
4.11	ROC and AUC for Dataset II for the single classifier, conventional bagging classifier and balanced bagging classifier	136
4.12	ROC and AUC for Dataset III for the single classifier, conventional bagging classifier and balanced bagging classifier	137
4.13	ROC and AUC for Dataset IV for the single classifier, conventional bagging classifier and balanced bagging classifier	138
5.1	Multimodal Arabic opinion mining framework	140
5.2	Multimodal Arabic sentiment analysis dataset creation process	144
5.3	Acoustic features extraction process	147
5.4	(a): Face detection phase, (b) Histogram of optical flow features extraction	149
5.5	Results for different investigated parameters for visual features	152
5.6	Single-level fusion techniques: (a) feature, (b) score, and (c) decision	153
5.7	Combined ORB and dense optical flow features extraction process	159
5.8	p-values of pairwise t-test of sentiment analysis system	161

6.1	Confusion matrix for demographic detection systems: (a) Gender, (b) Age-group, and (c) Dialect	168
6.2	p-values of pairwise t-test of demographics	169
6.3	Layout of the process for multi-class gender with sentiment detection .	170
6.4	Confusion matrix of multi-class demographic with SA: (a) Gender-sentiment, (b) Age-group-sentiment, and (c) Dialect-sentiment	172
6.5	Confusion matrix of multi-class gender, age-group and sentiment	173
6.6	p-values of pairwise t-test of multi-class demographics with sentiments	174
6.7	Confusion matrix for each label in multi-label system: (a) Sentiment only, (b) Gender only, and (c) Age-group only	176
6.8	Confusion matrix for pair of labels in multi-label system: (a) Sentiment-Gender, (b) Sentiment-Age, and (c) Gender-Age	177
6.9	Confusion matrix for tri-labels in multi-label system: Sentiment-Gender-Age	178
6.10	Per-class performance in terms of precision, recall and F_1 measure w.r.t.: (a) Single label, (b) Pair of labels, and (c) Tri-labels	178
6.11	Accuracy and average precision, recall and F_1 measure for Sentiment, Gender, Age, Sentiment-Gender, Sentiment-Age, Gender-Age, and Sentiment-Gender-Age	179
6.12	Compressions of multimodal gender, age and sentiment recognition at feature level fusion in case of HOF and HOF+HOG visual features . .	182
B.1	Significance Tests Taxonomy	195

LIST OF ABBREVIATIONS

ADASYN	Adaptive Synthetic Sampling Approach
ANN	Artificial Neural Network
ASTD	Arabic Sentiment Tweets Dataset
AUC	Area under ROC Curve
BoW	Bag-of-Words
CAE	Correlation-Attribute Evaluator
CART	Classification and Regression Tree
CBOW	Continoues Bag-of-Words
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DT	Decision Tree
FN	False Negative
FNR	False Negative Rate
FP	False Positive
GB	Gradient Boosting
GM	Geometric Mean

GMM	Gaussian Mixture Model
GNB	Gaussian Naïve Bayes
GRU	Gated Recurrent Unit
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradient
<i>idf</i>	Inverse Document Frequency
<i>k</i> -NN	<i>k</i> -Nearest Neighbors
LR	Logistic Regression
LSA	Latent Semantic Analysis
LSTM	Long Short Term Memory
MCC	Matthews Correlation Coefficient
MKL	Multiple Kernel Learning
MLP	Multi-Layer Perceptron
MSA	Modern Standard Arabic
NB	Naïve Bayes
NLP	Natural Language Processing
ORB	Oriented FAST and Rotated BRIEF
PCA	Principal Component Analysis
RBF	Radial Basis Function
RF	Random Forest
RNN	Recurrent Neural Network

ROC	Receiver Operating Characteristic
ROS	Random Oversampling
SA	Sentiment Analysis
SADAM	Sentiment Analysis Dataset for Arabic Multimodal
SE	Stacking-based Ensemble
SG	Skip-Grams
SMOTE	Synthetic Minority Over-sampling Technique
SVD	Singular Value Decomposition
SVM	Support Vector Machine
tf	Term Frequency
$tf-idf$	Term Frequency-Inverse Document Frequency
TN	True Negative
TNR	True Negative Rate
TP	True Positive
VE	Voting-based Ensemble
WEKA	Waikato Environment for Knowledge Analysis

DISSERTATION ABSTRACT

NAME: Saddam Hussein Mohammed Al-Azani

TITLE OF STUDY: Multimodal Arabic Sentiment Analysis for Video Opinion Mining

MAJOR FIELD: Computer Science

DATE OF DEGREE: May 2019

With the exponential rise of online content and social media such as blogs, twitter and TV talks, analyzing this multimodal content to identify the attitude or opinions of persons on certain topics or products has gained growing importance in a wide variety of applications, ranging from customer service, political and social debates, to cyber-security. Although a lot of work has been done for many languages to-date, especially English, not much has been done for the Arabic language. Moreover, it is expected that more effective sentiment and opinion analysis systems can be designed when combining different modalities of a person such as textual, audio and visual data. This study aims at exploring new approaches for automating multi-modal sentiment analysis and opinion mining for the Arabic language. We present a framework and develop a prototype that incorporates different modalities. The first multimodal sentiment analysis

database for Arabic is built. Several architectures and fusion methods are evaluated and a new methodology is proposed. This dissertation presents a new direction of sentiment analysis to detect speakers' gender, age-group and dialects with their sentiments. The experimental results demonstrate that demographic characteristics as features are able to detect sentiment with an accuracy rate of 74.63%. Incorporating demographic features with textual, audio and visual modalities has led to improving the performance in nearly all cases with achieving an accuracy rate of 94.66%.

ملخص الرسالة

الاسم: صدام حسين محمد العزاني

عنوان الدراسة: التحليل متعدد الأنماط للمشاعر والتنقيب عن الآراء العربية

التخصص: علوم الحاسب الآلي

تاريخ الدرجة العلمية: مايو 2019

مع تزايد وسائل الإعلام وشبكات التواصل المختلفة أصبح من السهل التعبير عن الآراء والمواقف تجاه موقف أو رأي أو منتج أو منظمة أو غيرها. مما أدى إلى الحاجة الملحة إلى تطوير أنظمة لتحليل المشاعر والتنقيب عن الآراء، وقد أثبتت أهميتها وفعاليتها من خلال انتشار تطبيقاتها في معظم المجالات المختلفة. أُجريت الكثير من الأبحاث في هذا المجال في مختلف اللغات الطبيعية وخاصة الإنجليزية إلا أن الأبحاث التي أُجريت على اللغة العربية محدودة ومازالت بحاجة إلى الكثير من الجهد والاهتمام. إن معظم الأبحاث التي أُجريت - حتى الآن - كانت لتطوير أنظمة لتحليل المشاعر والتنقيب عن الآراء في البيانات النصية. لكن مع تزايد انتشار الفيديو للتعبير عن المشاعر والآراء يمكن أن تكون هذه الأنظمة أكثر فعالية وكفاءة عند دمج عدة مصادر مختلفة للمحتوى كالصوت والصورة بالإضافة إلى التحليل النصي.

تهدف هذه الأطروحة إلى تقديم منهجية جديدة لتحليل المشاعر والتنقيب عن الآراء متعددة الأنماط (النص والصوت والصورة) باستخدام التقنيات الذكية المتقدمة. وقد تم أولاً الاستفادة من التقنيات والمصادر المتاحة لبناء الأنظمة الفردية لتحليل المشاعر والتنقيب عن الآراء من خلال النص والصوت والصورة. ثم قمنا باختبار دمج هذه الأنظمة الفردية للحصول على نظام متعدد الأنماط. ونظراً لعدم وجود قاعدة بيانات لهذا الغرض في مجال اللغة العربية، تشمل هذه الدراسة بناء قاعدة بيانات لتحليل المشاعر والتنقيب عن الآراء. وتم إجراء العديد من التجارب لاختبار كفاءة وقدرة النظام وتأثيره ومقارنة التقنيات المستخدمة والمقترحة. كما تناولت هذه الدراسة أيضاً التعرف على خصائص المستخدمين وتشمل الجنس والفئة العمرية واللهجة ومن ثم تقييم مدى ارتباطها بتحليل مشاعر المستخدمين. وقد أثبتت التجارب فعالية الطريقة المقترحة.

CHAPTER 1

INTRODUCTION

Sentiment Analysis (SA) is one of the most active research areas in Natural Language Processing (NLP) and is also widely studied in data mining, Web mining, and text mining [1, 2]. This field of study is important to the extent that it has spread to other sciences including management, politics, economics, and sociology. According to Liu [1], sentiment analysis is defined as “the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes”. Sentiment analysis is considered as the big umbrella for several tasks including: opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. As stated in the literature, the terms sentiment analysis and opinion mining can be used interchangeably.

There are many reasons behind the enormous interest of the research community in sentiment analysis. First of all, it has a wide range of applications, and is applicable in nearly every domain, such as branding and product analysis [3], expressive text-to-speech synthesis [4], question answering [5], analysis of political debates [6], tracking

sentiment timelines in online forums and news [7], conversation summarization [8] and health status analysis [9]. Second, there are still several gaps that have not been solved and many challenging research problems that need to be studied to build more reliable and effective systems. Third, sentiment analysis becomes a helpful and useful tool to analyze the rapid growth of user-generated contents which are expressed in several online media such as blogs, wikis, web forums and social networks. Through these platforms or environments, users give their opinions, post information, share knowledge, and get feedback from each others. Sentiment analysis is not a straightforward task and highly depends on the context and domain, so it is not always fixed. For example, the word “cheap” might be considered negative in politics but positive in economics.

A taxonomy of research work in sentiment analysis and opinion mining is shown in Figure 1.1. It can be classified in several dimensions based on tasks, approaches, granularity levels, and languages. Several tasks have been addressed including polarity determination (e.g. positive, negative), subjectivity detection, emotion recognition, sarcasm detection, aspect extraction, resource construction, and intention modeling. Various solution approaches have been investigated including supervised and unsupervised machine learning approaches, lexicon-based approaches, hybrid of machine learning and lexicon approaches, graph-based approaches and others. The analysis of sentiments or opinions have been performed at different textual granularity levels, including: document, sentence, word, aspect, concept, phrase, link-based clause, or sense levels [10]. Several researchers have worked on various single language sentiment analysis (e.g. English, Spanish, Chinese, etc.). Some have also considered

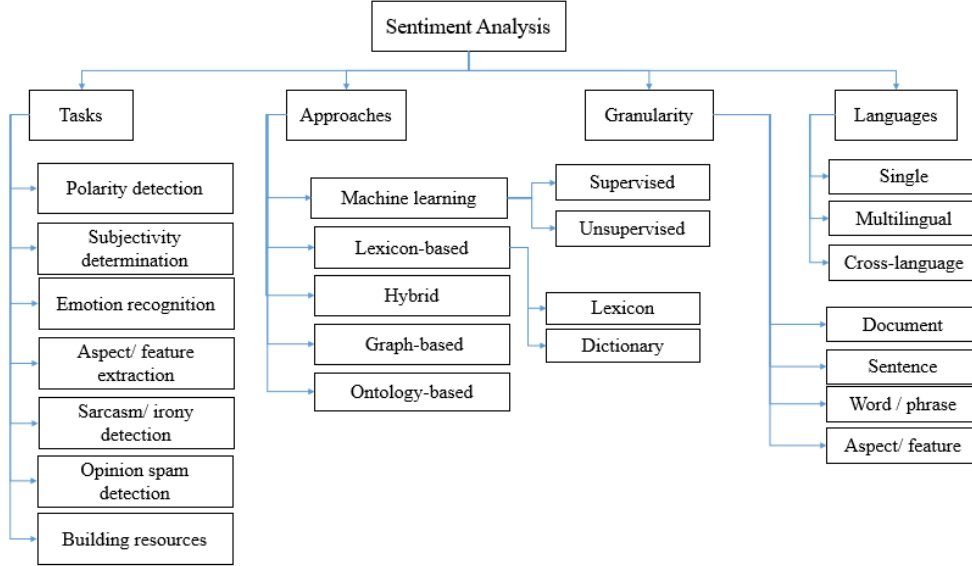


Figure 1.1: A taxonomy of sentiment analysis

multi-lingual and cross-language sentiment analysis. These approaches are reviewed in Chapter 2.

1.1 Problem Statement and Motivation

Sentiment analysis is important for analyzing online social media contents and products reviews, measuring economic indicators, etc. Most of the sentiment analysis studies have been concentrated on text-based analysis. Additionally, many available resources and corpora are compiled, developed and evaluated only on text modality. Nowadays, several social media platforms allow other forms of data to be used to express and represent people’s opinions including videos, audios and images. Thus, it is highly important to mine opinions and identify and fuse sentiments from the diverse modalities.

In this context, it is noteworthy to differentiate between two main concepts: *modal-*

ity and *medium* (plural. *media*). The first one “modality” is concerned with the sense by which a message is communicated between people or machines while the second concept “medium” focuses on the means of message communication [11]. Facial expressions and gestures are examples of modalities that can be sent and received via the same medium (video). The present coding scheme is concerned with modality rather than medium [12].

Sentiment analysis based on text modality is not a straightforward task and suffers from several issues related to morphological analysis, multi-dialects, ambiguity, temporal dependency, domain dependency, etc. In the same sense, recognition systems based on voice might be affected by different attributes such as low voice quality, background noise and disposition of voice-recording devices. This is true as well regarding visual modality, which can also suffer from illumination conditions, posture, cosmetics, resolution, etc. In consequence, this leads to inaccurate and insufficient representation of patterns. Multimodal information fusion aims at alleviating these issues. They provide several evidences for the same aspect which can lead to improving the performance significantly over the unimodal systems.

Most social media platforms were originated to share information in text-based format. Then, different types of data such as emojis, images, and audios were incorporated with texts. Recent studies indicate that social media platforms are pivoting from text to video content. The studies also stated that 80% of the shared contents will soon be videos because they have now taken over not only on YouTube but also on other social media platforms such as Facebook, Twitter, and Instagram. The contents of such videos review products, movies, visited places, healthcare, organizations, and

more. So, there is a need to propose, design and develop resources and tools to analyze and understand video contents; thus evolving from conventional unimodal analysis to more complex forms of multimodal analysis [13].

Detecting users' sentiment alongside with gender, age-group, dialect or/and nationality is very important and has several interesting applications. It is an excellent opportunity for large companies to capitalize on, by extracting user sentiment, suggestions, and complaints on their products from video reviews with their demographic characteristics. Consequently, they can improve, enhance their products/services to meet the needs of the customers. It can overcome the real-world gender, age and nationality bias issues of current sentiment analysis systems. For example, reviews of shaving machines by males are more significant than from females whereas reviews of women-specific products such as makeup products are more appreciated from females than from males. In addition, some products are specific for young people such as headphones and makeup. Reviewing such products by elder people are biased and results in wrong indicators for decision makers. Companies also might be more interested in reviews of citizens of some countries than others. Governments also need to explore issues related to the citizens according to their genders, age-groups, nationalities and dialects. This motivates us to propose a sentiment and demographic recognition approach as a new direction. A further important application of the proposed system is for adaptive and interactive educational systems. The content in the adaptive educational systems can be presented for beneficiaries according to their gender, age, dialect and emotion. The proposed system is, also, applicable in TV talk shows, video conference, and video messaging. Consequently, it supports decision

making in a wide spectrum of applications including product marketing, customer service, politics, healthcare, financial services, etc.

Arabic is one of the six most-spoken language with almost 422 million speakers [14]. Existing solutions to Arabic SA are limited compared to western languages approaches. The unique nature and complexity of the Arabic language requires researching and proposing appropriate solutions. Arabic alphabet is the second most widely used alphabet after Latin. It is a morphologically-rich language and has been classified into three categories based on its morphology syntax, and lexical combinations namely: Classic Arabic, Modern Standard Arabic (MSA), and dialectal Arabic.

1.2 Research Objectives

The main objectives of the proposed work include:

1. Conducting an intensive survey on the state-of-the-art methodologies and resources for automated sentiment analysis and opinion mining.
2. Providing a critical assessment of existing techniques for Arabic sentiment analysis.
3. Extending and building a comprehensive database for multimodal Arabic sentiment analysis and opinion mining; the existence of such database is very crucial for the advancement of research work in this area.
4. Investigating several structures and fusion methods and proposing a new methodology for multimodal Arabic sentiment analysis.

5. Developing a prototype for multimodal Arabic sentiment analysis and opinion mining.
6. Evaluating and benchmarking the proposed approach with related work.
7. Publishing and sharing the research outcomes and findings on this research topic with the research community for further advancement of this field.

1.3 Scope of Work

The focus of this dissertation is to conduct basic and applied research in the Arabic sentiment analysis and opinion mining utilizing different modalities. The scope of this work is as follows:

1. Research in the area of sentiment analysis is conducted using different modalities. It will lead to developing theory of sentiment analysis and producing algorithms and software tools/modules.
2. An Arabic multimodal sentiment analysis database is constructed. It will provide the research community with a benchmark to compare and contrast results from different systems and algorithms.
3. A prototype is developed to handle sentiment analysis from different modalities as a proof of concept.
4. Sharing the research findings with the research community through paper publication.

1.4 Research Design

The process of handling sentiment analysis, in this dissertation, is broadly broken down into several phases, as depicted in Figure 1.2, including: performing a comprehensive literature review, analyzing and utilizing the available resources, building the needed resources and/or extending the current resources, pre-processing, feature engineering, generating computational models, combining different modalities, and evaluation. To achieve the stated objectives the following methodology is followed:

1. **Comprehensive literature review:** Reviewing the related and recent published papers in reputable journals and conferences. Several bibliography databases are considered such as Scopus, Web of Science, and Google scholar, to find and analyze the literature. These papers are filtered and prominent related ones are selected. They are classified based on various attributes. Moreover, gaps and limitations of the existing approaches are highlighted.
2. **Analyzing existing resources:**
 - Analyzing the current resources including lexicons, datasets, corpora, pre-trained models, and utilizing them as main resources. Then extending the available resources to meet our research objectives and requirements.
 - Utilizing the current resources from the resource-rich language (English) to the focus language (Arabic) as needed.
 - Generating/constructing our resources including models and datasets.

3. **Multimodal sentiment analysis dataset construction:** Building a benchmarking multimodal sentiment analysis database from online social videos. The dataset will have some characteristics such as:

- Collected from online social videos.
- Being diverse in the sense that it is expressed by males and females of different ages, dialects, etc.
- Covering real-world conversations, i.e., not prepared in laboratory or in special environment or expressed by specific users.

4. **Preprocessing:** This dissertation deals with several modalities or information sources, including: text, visual (emojis), acoustic and visual (sequence of images). Each of them has its own preprocessing operations. For example, the main preprocessing operations for text include: punctuation marks removal, diacritical marks removal, stop words removal, noisy symbols removal, tokenization, normalization, etc. For visual modality, the face detection phase is considered to just detect the sentiment of speaker from his/her face. Then the detected phases are normalized and converted from BGR level to GRAY level. Each audio input is preprocessed as a ‘WAV’ format, 256 bit, 48000 Hz sampling frequency and a mono channel.

5. **Feature engineering:** In this phase, several features are investigated for each modality and feature extractors are developed. For textual features, different features are investigated including: Term Frequency-Inverse Document Frequency ($tf-idf$), Latent Semantic Analysis (LSA), structural features, and two

forms of word embedding features. Novel emojis based features are also proposed and evaluated. In addition, a combination of prosodic and spectral features are evaluated to represent acoustic features. Hybrid features of global and local descriptors are evaluated for visual features. Feature engineering task includes feature reduction and normalization.

6. **Classification:** This study evaluates several learning approaches including: shallow/base learning, ensemble leaning and deep learning.
7. **Unimodal development:** developing a unimodal approach for each source of information; that means we design and implement:
 - a. Textual sentiment analysis system,
 - b. Audio sentiment analysis system
 - c. Visual sentiment analysis system.
8. **Fusion:** Several fusion schemes and levels are evaluated. As shown in Figure 1.2, this phase integrates with feature engineering phase and classification phase. Feature-level, score-level and decision-level fusions are considered. Different methods are also proposed to combine these levels. The feature-level fusion relies on the feature engineering phase. Features are extracted from each content independently and then fused into a combined feature vector. Score-level and decision-level fusions rely on the classification phase. The features of each modality are extracted independently and then fed into a separate classifier. The classifier’s probabilities and decisions are fused at score level and decision level,

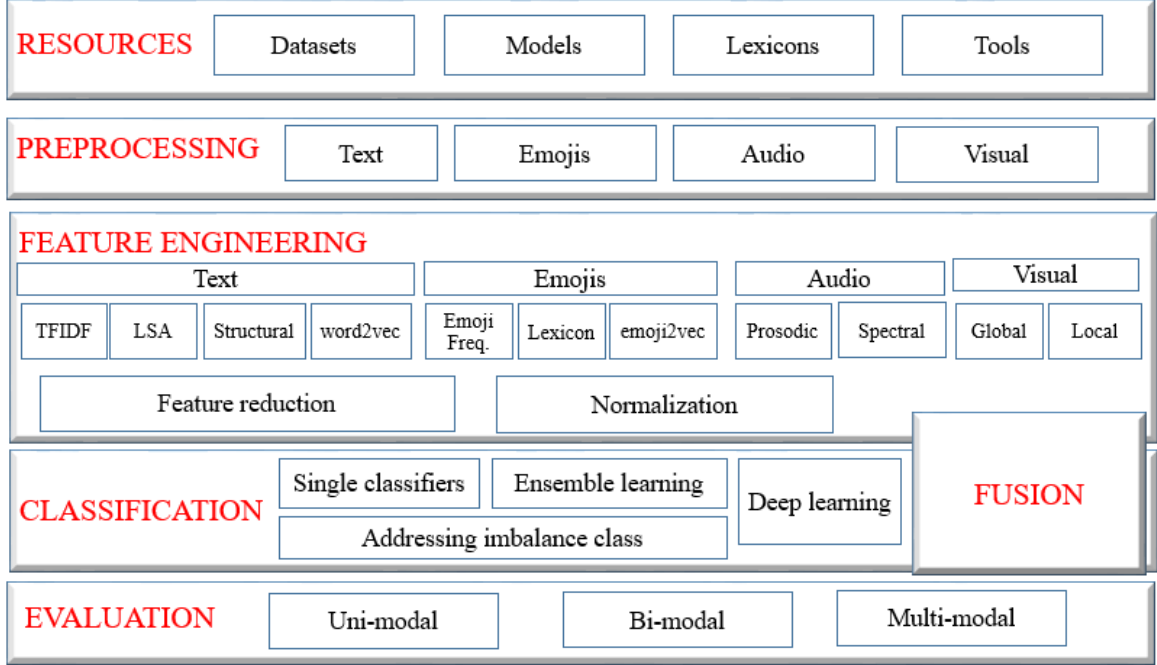


Figure 1.2: The high-level architecture of the methodology.

respectively. This results in several models representing single modalities, bi-modalities, tri-modalities (audio-text-visual) and quad-modalities (considering demographics as a new modality).

9. **Evaluation:** The generated models are evaluated using different evaluation methods and measurements.

1.5 Dissertation Contributions

This dissertation delivers several contributions including:

1. A comprehensive study is conducted to review the literature in terms of text-based approaches and multimodal sentiment analysis. The literature is classified

and compared according to some criteria. Several taxonomies and frameworks are proposed such as a framework to be followed by researchers for building multi-modal sentiment analysis datasets. A comprehensive literature review is also performed for emojis in social media and a taxonomy is provided according to emojis' applications, representations, issues, and approaches.

2. This dissertation presents a systematic empirical evaluation for text-based sentiment analysis. We first identify the popular methods and approaches for text-based SA and evaluate them then we address the related issues and limitations.

- *tf-idf*, LSA and structural features as traditional features and two forms of word embedding based features are utilized and evaluated to identify sentiments using publicly available datasets. We found that word embedding based features perform significantly better than traditional features. Consequently, word embedding based features are selected to be our main textual features in the remaining experiments of this study.
- The sentiment datasets are considerably imbalanced. Therefore, the class imbalance problem is addressed through evaluating different oversampling techniques with word embedding based features extracted from datasets with different imbalance ratios.
- Rare work and efforts have been conducted to utilize deep learning technology for sentiment analysis. This dissertation investigates various deep learning models based on Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) recurrent neural networks for sentiment anal-

ysis of Arabic microblogs. We designed and evaluated several deep learning architectures using CNN and LSTM along with word2vec word embedding technique. The proposed structures are evaluated using two publicly available Arabic tweets datasets. Promising results have been attained when combining LSTMs and compared favorably with most related work.

3. Nowadays, emojis are getting excessively popular in social media communication as a complementary way to quickly express opinions and ideas in a visual manner. This dissertation presents a novel approach for sentiment classification of microblogs based on non-verbal emoji-based features.

- Four feature extraction methods are proposed for emojis including two emoji embedding models, emojis frequencies, and emojis lexicon-based features. The effectiveness of the proposed approach on sentiment classification is analyzed in the context of the dialectal Arabic language using ten machine learning classifiers. The results are comparable to text-alone related features and even better than some of the traditional textual feature extraction methods.
- It integrates emojis with textual features to detect sentiments using several fusion levels and schemes as a bi-modal sentiment analysis problem. The experimental results reveal that emojis features can significantly improve the sentiment classification results when fused with text.
- It is found that, users tend to use emojis with positive polarity or happy emotion more than other polarities or emotions. This issue is considered

in this dissertation as a class imbalance problem and is addressed through generating synthetic instances for the negative opinions with the Bootstrap Aggregating (Bagging) algorithm. The performance is evaluated and compared on four datasets with a varying imbalance ratio ranging from two to more than 14.

4. The first systematic multimodal sentiment analysis of Arabic videos, based on text, audio and visual modalities is presented in this dissertation.

- It is based on three modalities: text, audio and visual.
- Different features are extracted to represent each modality. Prosodic and spectral acoustic features are extracted to represent the audio modality. Word embedding based features are adopted to represent the textual modality. For visual modality, dense optical flow descriptors are extracted. We also present a method for visual features extraction based on a combination of local and global descriptors.
- The considered modalities are combined in different fusion levels (feature, score and decision) with different schemes. The stand-alone modalities are combined at score level using SUM, Prod and MAX rules while the majority voting (MODE) rule is considered for fusion at the decision level. A multi-level fusion as a hybrid fusion method is presented and investigated.
- The experimental results illustrate that combining different modalities leads to a more accurate Arabic sentiment analysis system and improving the results of the standalone modalities significantly. The multi-level

fusion approach achieves the highest results.

5. We found that detecting sentiment alone will not satisfy the future requirements and needs. Sentiment analysis approaches might be biased to different users' demographics such as gender, age-group and dialect. In addition, companies need more information related to the detected sentiment such as the demographic information of users/customers. Combining gender, age-group, dialect or/and nationality with sentiment recognition is a more challenging problem for new business models and directed decision making. Therefore:

- This dissertation presents a multimodal approach to detect users' gender, age-group, dialect and nationality for Arabic speakers using audio, textual and visual modalities. The existing approaches are based on single modalities such as text, image, speech, or sequence of images (visual) individually. Different features for each modality are extracted and evaluated for the first time. Word embeddings are evaluated to detect gender, age, dialect and nationality of speakers. In addition, it applies a combination of prosodic and spectral features to detect those characteristics. We present visual features based on a combination of local and global descriptors.
- This dissertation presents a sentiment and demographics detection approach for Arabic speakers using audio, textual and visual modalities with reporting promising results.

6. This dissertation also analyzes the correlation between demographics and sentiment. The research question can be formulated as follows. Is users' demo-

graphic information capable to detect the sentiment? The experimental results demonstrate that demographic features are able to detect sentiment with an accuracy rate of 74.63%. This encourages us to consider demographic characteristic as a new modality to detect sentiment of Arabic videos. Incorporating demographic features with textual, audio visual modalities leads to improving the performance in nearly all cases. The highest results are obtained using four modalities.

7. This dissertation presents an ensemble neural network based fusion method to combine different modalities. Several structures of Artificial Neural Network (ANN) are presented and evaluated using standalone modalities. They are then combined using the proposed ensemble neural networks approach.
8. This dissertation provides and develops several resources, including:
 - A dataset for text sentiment analysis of Arabic microblogs in which each tweet contains at least one emoji. This is achieved through collecting several publicly available datasets and combining them, then filtering those instances that contain emojis. We found the resultant dataset is small of 1248 instances. Therefore, we increased it through collecting new subset of 843 instances and annotate it manually. We end up with a dataset of 2091 Arabic microblogs each of which contains at least one emoji.
 - Due to the unavailability of Arabic multimodal sentiment analysis dataset, a dataset is constructed. The dataset is collected from YouTube. It comprises of 63 opinion videos segmented into 524 opinions distributed as 250

negative and 274 positive utterances. The topics belongs to different domains including reviews of products, movies, cultural views, etc. The collected videos were recorded by users in real environments including houses, studios, offices, cars or outdoors with different settings.

- A dataset for multimodal demographic characteristics including: gender, age-group, dialect and nationality.
- Prototypes are developed to evaluate the effectiveness of the proposed systems.

9. Sharing the research findings with the research community through paper publication. The list of the dissertation’s publications is provided in Appendix A.

1.6 Dissertation Organization

In addition to this introductory chapter, the rest of the dissertation is composed of seven chapters. These chapters are as follows:

Chapter 2 provides a comprehensive literature review of different related approaches.

Chapter 3 presents sentiment analysis approaches based on textual based features. Different textual features are evaluated including hard-crafted based features and neural network language models. The class-imbalanced issue in sentiment analysis is, also, addressed in this chapter. Additionally, it investigates various CNN and LSTM deep learning models for sentiment analysis of Arabic microblogs. Parts of this chapter are published in [15, 16]

Chapter 4 presents the idea of adopting new non-verbal features for sentiment analysis of microblogs is explored. It presents different methods to extract features from emojis and build predictive models to detect sentiment. Several methods are investigated to combine emojis with texts in that chapter. Parts of this chapter are published in [17–20].

Chapter 5 describes the developed Arabic multimodal sentiment analysis dataset. A multimodal sentiment analysis approach is also presented. Different fusion techniques are evaluated to combine the different modalities. Parts of this chapter are published in [21–24].

Chapter 6 presents unimodal, bimodal and multimodal demographics recognition systems. It also presents a new direction to detect the sentiment of speakers along with their demographic characteristics. Furthermore, it investigates the correlation between users’ demographics and their sentiments. Parts of this chapter are published in [22–24] in joint with the previous chapter.

Chapter 7 concludes the work of this dissertation, and summarizes the main findings of each chapter. Possible future extensions are also discussed in that chapter.

CHAPTER 2

BACKGROUND AND LITERATURE REVIEW

This chapter provides a literature survey of the most related works. First, it reviews text-based approaches and classifies them based on different criteria. The different techniques applies and proposed to address imbalance class problem for sentiment analysis is presented. It also reviews deep learning based approaches for sentiment analysis. Opinion spam detection approaches are then reviewed and classified based on some different criteria. Existing multimodal sentiment analysis approaches are reviewed and categorized based on some attributes.

2.1 Text-based Sentiment Analysis

2.1.1 Text-based Sentiment Analysis Approaches

Sentiment analysis include several tasks, namely: subjectivity classification, polarity determination, emotion recognition, sarcasm detection, opinion words and aspects extraction, lexical and corpora creation (or building resources), entity recognition.

Subjectivity classification was addressed in [25–27], where the goal is classifying expressions as subjective or objective. As mentioned in [1], a sentence is objective if it expresses some factual information about the world whereas it is subjective sentence if it presents some personal feelings, views, or beliefs. An example of an objective sentence is “A new version of iPhone was released” while the sentence of “I dislike iPhone” is an example of the subjective sentence. The subjectivity in text can be further classified into positive, negative or neutral polarity. Polarity determination has been addressed by [28–39]. Emotion recognition is also considered a sentiment analysis task and is defined as “our subjective feelings and thoughts” [1, 40]. According to Plutchik [41], these emotions include: joy, sadness, anger, fear, trust, disgust, surprise, and anticipation. Emotion recognition was addressed in several studies such as [30].

One of the challenges of sentiment analysis is sarcasm detection. Sarcasm is often used to express negative feelings using positive literal expressions. Although, sarcasm detection is hard even for humans, it can contribute to improve the performance of many NLP tasks including sentiment analysis. For example, Hiai et al. [42] presented a rule-based method to extract sarcastic sentences in product reviews. The sarcastic sentences in product reviews were first analysed and categorized into eight classes by

focusing on evaluation expressions. The decision process consists of three phases. To evaluate this approach, they prepared two datasets: a development dataset which contains 34,917 sentences with 70 sarcastic sentences, and a test dataset which contains 33,864 sentences from 10,000 reviews. Their experimental results revealed the effectiveness of their method as compared to a baseline method. However, the precision rates of both methods were extremely low around 0.006 and 0.028.

Some researchers have addressed the task of feature and aspect extraction [43, 44]. The aspect in the sentence “The voice quality of this mobile is amazing,” is “voice quality” of the entity represented by “this mobile”. In this sentence the evaluation is only about the quality of voice and not the mobile in general. In contrast, the sentence “I like this mobile” indicates the aspect “general” of the entity represented by “this mobile”.

Building resources is a task that aims at creating lexicons, dictionaries and corpora for sentiment analysis and opinion mining. Providing a domain-specific lexicon might be an optimal choice rather than using the common lexicons for all domains. This is because that the polarity of a term changes from one domain to other so the polarity of a term might not be the correct for every domain. Several studies have been conducted to enrich the field of sentiment analysis and opinion mining by creating lexicons, datasets, tools, etc. Some of these studies are [32, 33, 38, 45–50].

Researchers also proposed another task that enriches the area of sentiment analysis and opinion mining by transferring a resource-rich language to another language rather than building or creating new resources from scratch. This task is called cross-language sentiment analysis. Mohammad et al. [51] evaluated two approaches: text translation,

and resource translation. In the first approach, the text of the focus language (e.g., Arabic) is translated into a resource-rich language (e.g., English). Then, a powerful sentiment analysis system of the resource-rich language is applied on the translated text. In the second approach, the sentiment-related resources of a resource-rich language, such as lexicons and labeled corpora, are translated into the focus language. Refaee and Rieser [52] translated MPQA English sentiment lexicon which was created and made publicly available by [53] and then manually filtered it to remove irrelevant or no-sentiment-bearing words. The resultant lexicon contains 2,627 entries.

Various approaches have been proposed to address the sentiment analysis tasks including: supervised and unsupervised machine learning-based, lexicon-based and hybrid approaches. Several machine learning techniques have been applied. SVM was applied in [33, 35, 36, 38, 43, 49, 51, 54]. Naïve Bayes (NB) approach was applied in [30, 33, 35, 36, 38, 44, 49, 54]. The k -NN approach was used in [33, 35, 36] and different algorithms of decision trees were applied in [33, 38, 49]. Logistic regression was used by [38, 49] and SGD in [38, 49]. Researchers also investigated advanced machine learning techniques such as ensemble classification methods in [25–27, 33, 34, 54]. An alternative model has been proposed to address sentiment analysis based on joining feature extraction and classification in a single integrated scheme. Such method is referred to as end-to-end learning, feature learning or deep learning model [39].

On the other hand, the lexicon-based approach is classified into dictionary-based approach and lexicon-based approach [1, 10, 40]. It depends on a list of common and precompiled entities to express positive or negative sentiments (sentiment lexicon or opinion lexicon). Entities might be words, phrases, idioms, lemmaADs, etc. Finding

sentiment polarities for the lexicon entities requires statistical or semantic methods. Liu [1] argued that “sentiment lexicon is necessary but not sufficient for sentiment analysis”. This is attributed to the fact that sentiment is domain- and application-dependent, which means that the positive sentiment word in a domain might be negative in another domain or application. For example, the word “cheap” is a positive sentiment in economics but negative in politics. Moreover, some sentences, such as interrogative sentences and conditional sentences, contain sentimental words but the context doesn’t express any sentiment. For example, the sentences “Can you tell me which Canon camera is good, please?” and “If I can find a good camera in the shop, I will buy it.” contain the sentimental word “good” but both sentences don’t express any positive or negative opinion on any specific camera. The first thing that one can do to solve this issue is to remove all conditional or interrogative sentences as a preprocessing step such as in [44]. This is a straight forward and trivial operation; however, and unfortunately, some of such sentences might express sentiments such as “If you need to buy a good camera, buy a canon” or “Does anyone know how to repair this terrible camera?” . It is difficult to deal with sarcastic sentences using this approach. Lexicon based approach have been applied in several studies including: [32, 42, 46–48, 50]. Hybrid approach was followed by using both machine learning techniques and lexicon-based approach in [31, 52].

The sentiment analysis and opinion mining has been conducted at different textual levels including: document, sentence, word, phrase, and aspect. Most of the reviewed studies, so far, have addressed the sentiments and opinions at the document level such as [25–27, 29, 29, 32–37, 47, 48, 50, 51, 54]. Sentence level analysis was considered

in [31, 39, 42, 44] while word level analysis was considered in [38, 49, 52] and aspect level in [43].

2.1.2 Imbalance Class Problem in SA

The imbalanced class problem has been addressed in several areas at the data level and/or algorithmic levels. Sampling techniques have been proposed to solve the imbalanced class problem at the data level to improve the predictive modeling capability. These techniques include oversampling, undersampling, and hybrid approaches. Oversampling techniques aim at balancing dataset through replicating or generating synthetic instances of the minority class. These technique vary in the way they generate synthetic instances. In contrary, the undersampling techniques aim at balancing the data distribution through eliminating instances of the majority class. However, eliminating instances randomly may result in eliminating important instances which may negatively affect the power of the generated models. Several methods have been proposed to overcome this issue such as NearMiss [55] and undersampling based on clustering [56]. An alternative direction is to address the class imbalance problem at the algorithm level, e.g. by improving the algorithm or using cost-sensitive learning, one-class learning, or ensemble learning [57]. Ensemble-based methods are powerful techniques to improve the classification performance [58]. However, individual classifiers need to be combined efficiently in order to obtain better results.

Hassan et al. [59] presented a bootstrap ensemble framework to alleviate class imbalance, sparsity, and representational richness issues. Experimental results revealed that this approach can lead to more accurate predictions across sentiment classes, as

compared to other considered tools and algorithms.

Ah-Pine and Morales [60] used oversampling techniques: Synthetic Minority Over-sampling Technique (SMOTE), Borderline-SMOTE and Adaptive Synthetic Sampling Approach (ADASYN) to address the class imbalance problem with textual based features. They carried out the experiments on three imbalanced datasets expressed in English and French. The decision tree CART algorithm and logistic regression based classifiers were used for evaluation. It was reported that oversampling techniques are able to reduce the bias towards the majority and improve the recognition of the minority class as well as the geometric mean criterion.

An approach presented in [61] to analyze products reviews. This approach relied on modifying the data distribution and the classifier to alleviate the issue of class imbalance. A modified version of the bagging classifier is used where the dataset was sampled into sub-samples which are used to train different base learners. Instead of creating subsets using bootstrap sampling, consecutive important subsets of instances were constructed using the modified bagging approach. An instance was considered to be important if it improves the diversity. Using unigrams, bigrams, and trigrams textual features, the proposed approach is compared with Support Vector Machine (SVM) and classical bagging and the reported results showed its superiority over the other approaches in terms of the area under the receiver operating characteristic curve (AUC-ROC).

A method referred to as iSRD is presented in [62] to address the imbalance class issue of spam review detection. It is based on generating several balanced subsets by under-sampling the majority class. Then a classifier is trained from all sampled

datasets to create an ensemble for spam classification. It was reported that iSRD significantly performs better than C4.5 in terms of True Negative Rate (TNR), False Negative Rate (FNR), Sensitivity, Precision and Area under ROC Curve (AUC).

In the study of [63] a semi-supervised learning method was presented for sentiment classification based on random subspace generation. The undersampling approach was utilized to generate several random subsets of balanced initial training data. The generated subspaces can train ensemble classifiers to select confident instances from the unlabeled data, in the same way as co-training. To make variation among the involved classifiers, several different subspaces were created dynamically in an iterative manner. It was reported that the presented method can utilize the unlabeled data successfully and performed better than the static subspace generation. Particularly, the undersampling approach randomly selects several subsets of the majority class instances from the initial training data and then combines them with all the minority class instances to form new training sets. Given a new balanced training data, any existing semi-supervised learning method can be applied to use the unlabeled instances.

For Arabic imbalanced datasets, Mountassir et al. [64] addressed the sentiment analysis and presented three undersampling methods, namely Remove Similar, Remove Farthest and Remove by Clustering. These methods were evaluated using Naïve Bayes, SVM and k -Nearest Neighbors (k -NN). In two cases out of six, the Remove Farthest method performed better than the random undersampling method in terms of Geometric Mean (GM) score.

Refaee [65] employed SMOTE by experimenting on an imbalanced dataset of 6,894 Arabic tweets. Word unigrams and bi-grams are used as features with an SVM clas-

sifier. It is reported that applying SMOTE significantly improved results in terms of F_1 and accuracy. However, the results were significantly degraded when applying SMOTE with a larger dataset. This might be attributed to that the large dataset is composed of a large number of features and is annotated automatically (so the expected noise is much higher).

2.1.3 Deep Learning based Approaches for SA

Recently, due to the remarkable success of deep learning in computer vision, it has been attempted for other domains including natural language processing. Deep neural language models has been successfully applied for feature extraction. The main advantage of these models is that they don't require any feature engineering for learning continuous text representation from data. Instead, deep contextual features about words are extracted in a lower dimensional space. Many techniques have been proposed for learning word vectors such as word2vec [66, 67]. Other deep learning models that have been applied to NLP including CNN [68–70] and LSTM [71]. For instance, Kalchbrenner et al. [69] introduced a dynamic CNN for modeling sentences and evaluated it for sentiment prediction and question classification demonstrating good performance. Kim [68] presented an improved scheme based on CNN which employs dynamic and static word embeddings simultaneously for sentence classification and evaluated it on English sentiment analysis.

In supervised machine learning approaches, the studies conducted on sentiment analysis can be classified into four different directions. The first direction applies hand-crafted features to train traditional or shallow classifiers such as SVM, Multi-Layer

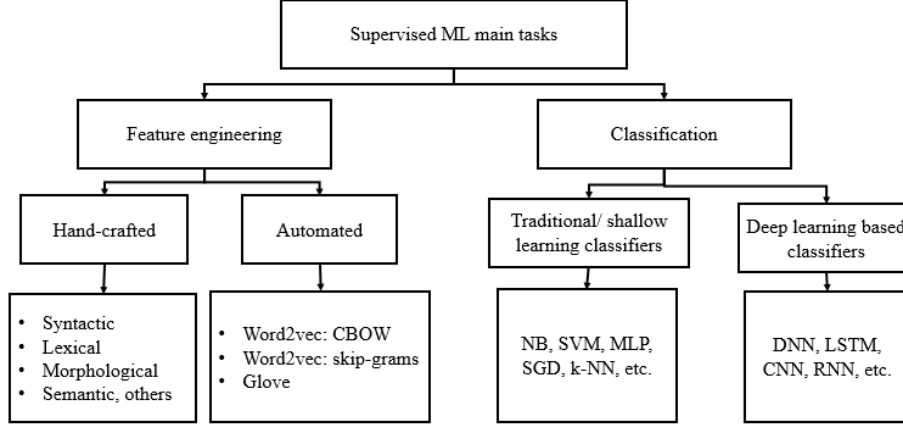


Figure 2.1: Type of machine supervised learning approaches

Perceptron (MLP), NB, and decision tree classifiers. The second direction applies hand-crafted features with deep learning classification methods such as Deep Neural Network (DNN), Recurrent Neural Network (RNN), LSTM and CNN. Examples of research work belonging to this direction are those conducted in [39, 72–74]. The third direction applies automated features generated by word embedding techniques such as word2vec [67, 75] and GloVe [76] with traditional or shallow classification methods. An example of this group the study conducted by [77]. The final direction applies automated features extraction methods with deep learning classification methods, e.g [68, 69, 73, 78]. We refer to this type as pure deep learning sentiment analysis. Fig 2.1 shows a taxonomy of the machine-learning methods applied to sentiment analysis.

Liu et al. [72] presented a hybrid method for bilingual text sentiment via integrating deep and shallow learning features. LSTM, NB-SVM, word vectors and bag-of-words were utilized. The approach was evaluated using dataset of NLPCC 2014 for binary class polarity detection of reviews in English and Chinese.

Alayba et al. [73] addressed Arabic sentiment analysis on health services. NB, SVM and logistic regression were applied and trained using n-grams and *tf-idf* features. In addition, deep neural networks learnt using word features was used. Moreover, they also used CNN with word2vec based features. The best results reported using SVM.

Al-Sallab et al. [39] addressed the problem of Arabic polarity detection using different deep learning architectures. They considered deep belief networks, deep auto encoders trained using Bag-of-Words (BoW) features. In addition, they used recursive auto encoder to cope with the lack of context handling in the deep belief networks, deep auto encoders architectures.

Abbes et al. [74] applied DNN and LSTM to detect polarity in Arabic reviews. A set of 1800 book reviews expressed in modern standard Arabic was extracted from LABR dataset to test the effectiveness of the used models. TF-IDF features were used to learn the classification methods. The highest results in terms of accuracy, precision, recall and F1 were reported using LSTM.

Altowayan and Tao [77] used Continoues Bag-of-Words (CBOW) based features to learn several classifiers including SVM, decision trees, NB, random forests for subjectivity and polarity detection. Different datasets are used to evaluate the proposed method expressed in modern standard Arabic and dialectal Arabic with different genres.

Kalchbrenner et al. [69] proposed a dynamic CNN for modeling sentences and tested its effectiveness for sentiment prediction and question classification and high results were reported. Kim [68] proposed an improved method based on CNN which employs dynamic and static word embeddings simultaneously for sentence classifica-

tion. The effectiveness of the proposed scheme was tested using English sentiment analysis.

Dahou et al. [78] presented a pure deep learning method for Arabic polarity detection of tweets and reviews. They first evaluated CBOW and skip-grams techniques and found that CBOW more efficient than skip-gram. Accordingly, they applied CBOW with a CNN architecture similar to [68].

2.2 Opinion Spam Detection Approaches

Opinion spam detection is a task of sentiment analysis [79]. It aims at detecting automatically spam opinions using techniques that usually depend on content of the review, review meta-data, and real-life knowledge about the reviewed entity. An opinion spam is recognized as false reviews expressed to promote a low quality entities (e.g., products, services, individuals, etc.) using positive opinions or to damage the reputation of a given entity with negative opinions.

Spambots are malicious computer programs well-known for sending junk or unsolicited emails over the Internet, which can endanger email users. Recently, spam has become a rising critical problem in social networking services such as Twitter with huge volume of fake posts everyday. In 2010, it was reported that there are around 5% spam tweets in Twitter [80] while in 2017, from 9% to 15% of Twitter accounts are estimated to be controlled by software (bots) [81]. Creating and publishing automatically-generated tweets that occupy a good portion of the continuous stream of tweets can be easily abused to hinder information extraction applications [82, 83].

Different view of points are assumed to consider a tweet or a text as an spam. For example the study of [84] considered that spammers post tweets contain typical words of a trending topic and URLs, usually obfuscated by URL shorteners, that lead users to completely unrelated websites. This kind of spam can contribute to de-value real time search services unless mechanisms to fight and stop spammers can be found.

The tweet is considered as an spam in [85] if it just contains a hashtag, a mention, a URL or an image without purely text and it is out of context, e.g. the tweet's content or sentiment is irrelevant to the context in which the tweet is embedded. They also considered consider any tweet advertising a paid retweet/favorite service or selling followers to be a spam as well. On the other hand, Almerkhi and Elsayed [83] to determine whether a tweet is generated by human (manually generated tweets) or by bots (automatic generated tweets) claimed that "automated" tweets, might be partially-edited by a human, or completely automated (e.g., prayer times or temperature readings). Arabic bots often use formal or MSA in their messages. Furthermore, tweets generated by bots are not personalized, as they discuss broad topics like news and famous quotes.

Non-legitimate tweets in [86] were considered as either fake or spam. It is a fake if it satisfies at least one of the following conditions: (1) incorrect location related to the event (2) incorrect time/date related to the event (3) some other incorrect information related to the event (4) link to misleading/ fake image. On the other hand they also in the same study defined a tweet as spam if it satisfies at least one of the following conditions: (1) link to a spam page (pharmacy, loans, etc) (2) link to a pornographic content (3) link to advertisements (personal agendas, etc). If a given tweet is neither

fake nor spam it is considered as legitimate tweet. Opinion spam detection approaches based on the language are classified in the following subsections.

2.2.1 Non-Arabic Opinion Spam Detection

The problem of detecting spammers on Twitter is considered in [84]. A large dataset of almost 1.8 billion tweets contain 1.9 billion links expressed by more than 54 million users was first collected. A large labeled collection of users , manually classified into spammers and non-spammers using tweets related to three famous trending topics from 2009. It contains 8,207 users distributed as 355 spammers and 7,852 non-spammer. They deal with the imbalanced problem by selecting a set of 710 non-spammers randomly which is twice of the number of spammers. A set of user/profile-based features and tweet content-based features were extracted to detect spammers. It was reported that around 70% of spammers and 96% of non-spammers were correctly classified.

Alberto et al. [87] developed an online system called, TubeSpam, to filter comments posted on YouTube. First, they evaluated several classifiers for YouTube comment spam detection. It is reported that the statistical analysis of results indicate that, with 99.9% of confidence level, decision trees, logistic regression, Bernoulli NB, random forests, linear and Gaussian SVM are statistically equivalent. Five datasets composed by real, public and non-encoded data were collected from YouTube through its API. They contain 1956 comments labeled manually as 1005 spam and 951 legitimate comments (almost balanced). Content based features, particularly, term frequency are used as features.

Rajdev and Lee [86] performed a case study of 2013 Moore Tornado and Hurricane Sandy. They presented flat and hierarchical classification approaches to detect both fake and spam tweets and distinguishing between them. Our experimental results show that our proposed approaches identify spam and fake messages with 96.43. They randomly selected 1,050 out of 9,284 tweets relevant to 2013 Tornado dataset consisted of 350 non-legitimate (i.e., 21 fake tweets and 329 spam tweets) and 700 legitimate tweets

Wang [88] presented an approach to detect the spam bots from normal ones for tweets. Both user/profile-based features and content based features are extracted. Three graph-based features, such as the number of friends and the number of followers, are extracted as user based features to explore the unique follower and friend relationships among users on Twitter. Three content-based features are also extracted. Several classifiers are applied namely: decision tree, neural network, SVM, and k-nearest neighbors, to identify spam bots on Twitter. Bayesian classifier outperforms others. A dataset of 25,847 users, around 500K tweets, and around 49M follower/friend relationships were collected. 500 Twitter user accounts were annotated manually to two classes: spam and not spam by reading the 20 most recent posted tweets per each user and checking the users' friends and followers. The result shows that there is around 1% spam account in the data set. To mitigate the imbalanced problem, the minor class is over-sampled manually by adding more spam data.

Wang et al. [89] presented a general spam detection framework to be used across all social network platforms. once a new type of spam is detected on one network, it can automatically be identified on the other networks as well. A profile model was defined

using 74 attributes and a message model was defined using 15 common attributes used in messages such as “To”, “From”, “Timestamp”, “Subject”, and “Content”. In addition, they defined a web page model based on common HTTP session header information.

A large dataset of over 600 million tweets was collected by [90]. Almost 6.5 million spam tweets were annotated and 12 lightweight features related to user/profile and content were extracted to be utilized for online spam detection. Several experiments were carried out using six machine learning classifiers: Random Forest, C4.5 Decision Tree, Bayes Network, NB, k -NN and SVM. They were applied under various conditions to evaluate their effectiveness and weakness for timely Twitter spam detection.

2.2.2 Arabic Opinion Spam Detection

In this part the focus on research addressed spam detection on Arabic language. El-Mawass and Alaboodi [85] presented a method for detecting accounts that promote spam and content pollution on Arabic Twitter. The spam content on Saudi Twitter was analyzed using the state-of-art features on a large crawled dataset of more than 23 million Arabic tweets, and a manually labeled sample of more than 5000 tweets. They also adapt the previously proposed features to respond to spammers evading techniques, and use these features to build a new highly accurate data-driven detection system. Several features are extracted related to profile and content to classify tweets as spammer or non-spammer. NB, Random Forests and SVM with Radial Basis Function (RBF) kernel implemented on Waikato Environment for Knowledge Analysis (WEKA) are used. Several metrics are used for evaluations.

Almerekhi and Elsayed [83] presented an study to detect whether a tweet is generated automatically (e.g., by bots) or manually (by human). They used formality, structural, tweet-specific, and temporal features over about 3.5 k randomly sampled Arabic tweets. It was reported that classification based on the aforementioned features outperform the baseline unigram-based classifier in terms of classification accuracy. Additionally, combining tweet-specific and unigram features improved classification accuracy to 92%, which is a significant improvement over the baseline classifier, constituting a very strong reference baseline for future studies. experimented with three classification algorithms: SVM, NB, and Decision Trees implemented on WEKA. Two sets of Arabic tweets were created: the first one includes 1.2-million tweets (represented by their tweet ids). The second contains a total of 3503 manually-labeled tweets, where 1944 were labeled as automated tweets and 1559 were labeled as manual tweets.

Mataoui et al. [91] presented an Arabic content spam detection system based on a set of both profile/user and content-based features which characterize Arabic spam content. The dataset was posts and comments collected from Facebook platforms. A set of 9697 comments contain 1112 spam and 8585 non-spam were collected. The issue of imbalanced class problem was addressed manually such that 7473 non-spam comments were removed randomly. Several profile/user- and content-based features are used, namely: Comment size, Number of lines, Number of hashtags, Number of emoticons, Number of diacritics, Existence of specific sequences, User publication frequency, Repetition frequency of a comment, Similarity between post and comment topics. Seven classifiers , namely: NB, J48, SMO, Decision Table, Logistic Regression

Table 2.1: Opinion spam detection approaches

Ref	Problem	Features		Reduction	Dataset			Addressing balancing	Classifiers	Lang.
		content	user		balanced	labelling	source			
[84]	spammers and non-spammers	✓	✓	✓	×	Manually	Twitter	undersampling randomly	SVM	English
[87]	spam & legitimate	✓	×	-	✓	Manually	YouTube	NA	Several	English
[86]	legitimate&fake & spam	✓	✓	-	×	Manually	Twitter	×	Several	English
[88]	spam & legitimate	✓	✓	-	×	Manually	Twitter	oversampling manually	several	English
[89]	spammers and non-spammers	✓	✓	-	×	Manually	cross social platforms	×	Several	English
[90]	spammers& non-spammers	✓	✓	-	×	Manually	Twitter	undersampling manually	several	English
[85]	spammers and non-spammers	✓	✓	-	×	Manually	Twitter	×	Several	Arabic
[83]	automated vs. manual tweets	✓	✓	-	×	Manually	Twitter	×	Several	Arabic
[91]	spam & legitimate	✓	✓	-	×	Manually	Facebook	undersampling randomly	Several	Arabic

Classifier, SGD implement on WEKA.

Table 2.1 presents a summary of the reviewed related works and compare them based on some attributes namely: addressed problem, type of features, reduction techniques, description of dataset (is it balanced? how does it annotated? what is the source?), addressing balancing method, used classifier and the language. It is clear that, all research reviewed in this study either deal with the imbalance dataset or balancing the used dataset manually either by collecting more examples for the minority class or removing examples from the majority class.

Opinion spam detection problem is naturally highly imbalance class problem. However, the reviewed studies either address this issue manually through undersampling the major class or ignore it. This motives us to explore the impact of imbalance ratio on the performance of Twitter spam detection using multiple approaches of single and ensemble classifiers. Besides ensemble-based learning (Bagging and Random forest), we apply the SMOTE oversampling technique to improve detection performance especially for classifiers sensitive to imbalanced datasets. Applying the oversampling technique significantly improved the results in most cases, especially for SVM-based classifiers. This study is published in [92]. Another finding is that none of the reviewed studies has evaluated word embedding based features to detect spam opinions. This

also motivates us to explore word embedding techniques as textual features to detect spam in Arabic tweets on a dataset of 3503 instances prepared by [83]. Three machine learning classifiers were used to evaluate the proposed features including: NB, Decision Tree (DT) and SVM. The experimental results reveals that:

- Word embedding techniques are able to detect Arabic spam tweets with 87.32% accuracy, 87.40% precision, 87.33% recall and 87.33% F_1 .
- Models generated using skip-gram are more efficient than those generated using CBOW in most cases.
- Models generated using Twitter domain outperform other text domains used to learn word embedding models.
- SVM classifier outperforms other classifiers significantly.

This study has been accepted and presented and will be published soon in [93].

2.3 Multimodal Sentiment Analysis

Most of the works on sentiment analysis have been performed on text-based sentiment analysis. Even most of the available resources and corpora are designed, evaluated and compiled for text-based sentiment analysis only. Nowadays, social media platforms are allows users to use multimedia (text, images, audio and video) to represent their opinions. Thus, it is highly important to mine opinions and identify sentiments from diverse modalities. So far, the field of multimodal sentiment analysis has not received much attention. This section presents a comprehensive review of recent studies that

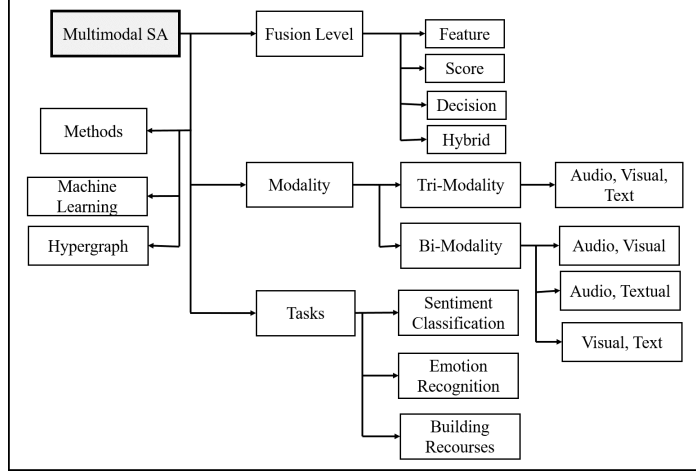


Figure 2.2: Multimodal sentiment analysis taxonomy

have two or modalities. A taxonomy is presented for the multimodal sentiment analysis shown in Fig. 2.2 and Table 2.2.

Table 2.2: Multimodality and tasks

Modality	Sentiment Classification	Emotion Recognition	Sentiment Classification and Emotion Recognition	Building Resources
Audio, Visual, Textual	[94–102]	[103–105]	[106]	[94, 95]
Visual, Textual	[107–113]		[114]	[111, 112, 114]
Audio, Textual	[115]	[116, 117]		
Audio, Visual		[118–120]		[118]

2.3.1 Fusion Approaches

There are various directions to integrate multiple modalities depending on which modalities are chosen and at which level they are fused. As shown in Figure 2.3, there are three modalities: Textual (T), Audio (A), and Visual (V). This gives as four possibilities to combine them, which are: A-T, T-V, A-V, and V-A-T. The fusion can be performed at feature level, score level, decision level or hybrid. Figure 2.4 illustrates the fusion levels of tri-modality of sentiment analysis at feature level, score

level or decision level.

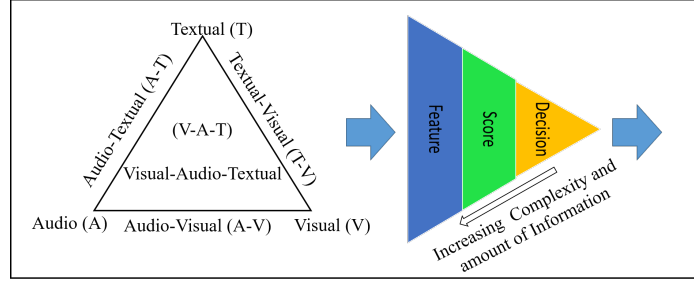


Figure 2.3: Fusion Levels

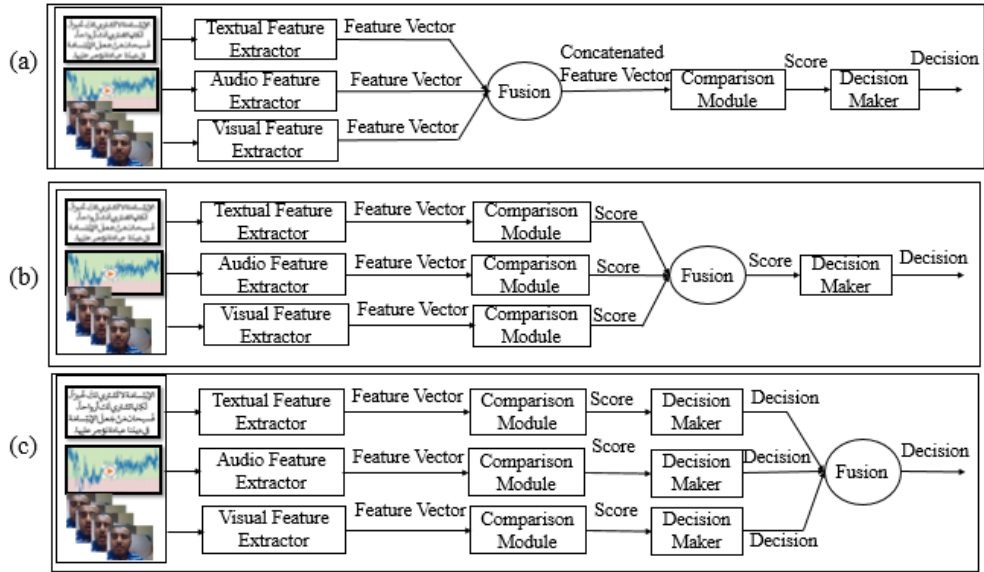


Figure 2.4: A tri-modal sentiment analysis system showing the fusion at: (a) feature level, (b) score level and (c) decision level

Feature level fusion is also known as early fusion in which the extracted features for each modality are combined using a certain strategy to generate a new feature vector. The generated vector is a high-dimensional data and might result in the curse of dimensionality problem. Variations of input data is another issue related to this level of fusion since the features are extracted from different channels and represented in different scales. It often requires different process when combined such as normalization,

Table 2.3: Fusion Levels

Reference	Feature	Score	Decision	Hybrid
[94, 96, 103–108, 111–115, 118]	✓			
[109, 116, 117]			✓	
[95]				✓
[98, 99, 110, 120]	✓		✓	
[119]	✓	✓		

transformation and reduction. There are several normalization schemes including: min-max normalization, Z-score normalization, Tanh-estimators normalization, etc. Feature level fusion provides much information so it is more flexible than other levels. Another form of feature level fusion is by combining features extracted from different feature extractors for the same modality. Feature level fusion is implemented in most studies for multimodal sentiment analysis.

In fusion at score level, the scores of the similarity between the input and template feature vectors of each modality are combined. Score level fusion is simple for implementation and provide much information than the decision level fusion. The combination is performed using different methods including: density based score fusion, transformation based score fusion, and classifier based score fusion. Similar to feature level fusion, score level fusion requires performing score normalization because the original scores obtained from different modalities and represented in different scales.

Decision level fusion is also known as late fusion and in which the final decision is made based on the local decisions of each individual modality. This level is easy to be implemented but it is often computationally expensive due to the various classification methods.

The hybrid fusion is conducted by utilizing and combining aforementioned fusion levels. It might include the advantages of other fusion levels when it is developed perfectly. A study of [95] applied the hybrid fusion level such that they combined the audio and visual modalities using the feature level fusion and the resultant model is combined with the textual modality using the decision level fusion.

2.3.2 A-T Sentiment Analysis

Govindaraj and Gopalakrishnan [115] fused audio and textual modality to perform an intensified sentiment analysis on customer product reviews. A set of acoustic features and a set of lexical features were extracted and then combined. SVM was used to predict the sentiments of customers. This method was evaluated using an audio speech dataset downloaded from YouTube. It was prepared from Amazon product reviews. It is reported that this bimodal approach performed better than each modality. The accuracy rate obtained using the textual and acoustic models individually were 70.50% and 65.62%, respectively. However, the accuracy rate obtained using the bimodal approach was 83.33%.

Wu and Ling [116] presented emotion detection approach of affective speech using acoustic-prosodic information and semantic labels. For aucostic-prosodic detection-spectrum, formant and pitch-related features were extracted as well as an ensemble classifier of Gaussian Mixture Model (GMM), SVM and MLP based on meta decision tree was used. Semantic labels derived from HowNet, a Chinese knowledge base, were used to extract emotion association rules automatically. The experimental results illustrated that applying ensemble classifier achieved higher accuracy than the

individual ones. Additionally, combining acoustic and textual modalities resulted in improving the performance. However, the dataset collected for this study is composed of 2,033 sentences collected in a lab environment.

Abburi et al. [117] presented a multimodal sentiment analysis approach from textual and audio information to detect the sentiment of Telugu songs. The textual lyric features were extracted from the bag of words and Doc2Vec generated a single vector for each song. Several supervised machine learning algorithms were used to classify sentiments of the textual features including SVM, NB and an ensemble of both classifiers. On the other hand, audio features were utilized as an add-on to the lyrical ones. The audio features include prosody features, temporal features, spectral features, tempo and chroma features. Different classifiers were also to predict the sentiment using the audio features including GMM, SVM and the ensemble of GMM and SVM. The approach was evaluated on Telugu database which was collected from the YouTube. It contains of 300 Telugu movie songs and lyrics corresponding to each song which was annotated as Happy and Sad. Combining the two models resulted in improving the performance of sentiment analysis. These text and audio features are extracted at the beginning of the song, at the end of the song and for the whole song. It was reported that the performance obtained from the first 30 seconds of a song outperformed that was obtained from the last 30s or from the whole song.

2.3.3 T-V Sentiment Analysis

Borth et al. [114] presented a method to create a large scale visual sentiment ontology (VSO) automatically. VSO is composed of 3000 Adjective Noun Pairs (ANP).

They introduced a new visual concept detector library called SentiBank to detect the presence of 1200 Adjective Noun Pairs in an image. For visual based sentiment analysis, they proposed SentiBank mid-level representations and compared them with low-level features including color histogram, GIST, LBP, and BoW. Two classifiers were used for this purpose: Linear SVM and Logistic regression. It was reported that mid-level features performed much better than low-level features. Additionally, it was reported that the multimodal approach outperformed the individual ones and Logistic regression outperformed the Linear SVM.

Chen et al. [107] presented a multimodal hypergraph-based method to determine microblog sentiment from textual, visual and emoticon information. The hypergraph structure captures the similarities of tweets on different modalities. Each vertex, in the constructed graph, represents a tweet and the hyperedge was formed by the “centroid” vertex and its k-nearest neighbors on each modality. To learn the relevance score among tweets, the transductive inference was conducted. This approach was evaluated using a dataset of over 6,000 microblog tweets collected from Sina Weibo. The accuracy of 86.77% was reported with 7% improved compared to the literature. It was reported that this method performed better than naïve Bayes, SVM and logistic regression.

Baocchi et al. [108] presented a unified model of both textual and visual information for sentiment analysis of micro-blogging content. The proposed method to predict the polarity of sentiments works. To obtain the textual representation, the continuous Bag-of-Word (CBOW) features were extracted for one tweet text window at a time. The associated image of the tweet was represented by extracting features

using denoising autoencoder (DA). The textual and the visual representations were combined (feature level fusion) and a polarity score is obtained using the logistic regression. Each window polarity was summed into a final tweet polarity score. LR was also compared with SVM and outperformed SVM. Four datasets were used in this work namely: Sanders Corpus, Sentiment140 [121], SemEval-2013 and SentiBank Twitter Dataset. The first three datasets are textual data compiled from Twitter while the fourth one (SentiBank) is composed of 470 positive tweets with accompanied images and 133 negative tweets with accompanied images (total 603 tweets). The dataset is related to several topics (21 topics) and labeled using Mechanical Turk. It was reported that the combined modality outperformed the individual modality. The highest accuracy rate of 79% was reported for SentiBank dataset which is higher than the accuracy rate, 72%, reported in [114].

Yu et al. [109] presented a multimodal framework from both textual and image content by using deep learning in a CNN to analyze the sentiment in Chinese microblogs. Firstly, CNN was trained on top of pre-trained word vectors for textual sentiment analysis and another deep convolutional neural network was employed with generalized dropout for visual sentiment analysis. The proposed framework was then tested using a Sina Weibo dataset. It was compiled by from Sina-Weibo, Chinese social media network and contains text and related images. It covers several topics including weather events such as typhoons and smog, products such as iOS7 and Meizu MX3, and gossip about celebrities and films. It was reported that the multimodal method performed better than the textual or visual modality individually.

You et al. [110] presented a cross-modality consistent regression (CCR) model

to enforce the agreement between sentiment labels predicted by visual and textual modality features. CNN was fine-tuned on images collected from Getty Image and a paragraph vector model was trained on the related titles and descriptions of the images to learn textual features. A total of 588221 weakly labeled images with their titles and descriptions were collected from Getty Images and labeled based on a list of 101 sentiment keywords. Additionally, 31,584 tweets of images and English text were collected and filtered as another test dataset. They were weakly labeled using VADER [122]. Several experiments were performed using machine weakly labeled and manually labeled image tweets. It was reported that the CCR model improved the performance than the visual modality and textual modality individually as well as the early and late fusion methods.

Zadeh et al. [111] presented a multimodal sentiment analysis using verbal and visual data. As for verbal features a set of bag-of-words from monograms and bi-grams was created from words in speech segments with including speech pauses and pause fillers. However, facial gestures were used as visual features. Facial gestures include smile, frown, head nod, and head shake. Support vector regression (SVR) was used. It was reported that the combination of verbal and visual model performed better than the individual ones. Additionally, the authors created a multimodal dictionary includes a simple representation of words and gestures. The entities of the dictionary was represented as a product of words and gestures (w, g) where $g \in G$ and $G = \{\text{smile, frown, head nod, head shake, } \sim\text{smile, } \sim\text{frown, } \sim\text{head nod, } \sim\text{head shake}\}$. It was also reported that the multimodal dictionary model performed better than the combination of verbal and visual model such that mean absolute error (MAE) of 1.10

and correlation of 0.53 were reported.

Kang et al. [112] presented a multimodal sentiment analysis to identify the users with depressive moods by exploring images, emoticons and texts. This is performed by analyzing the daily tweets of the users for a long period of time single modal analysis were first performed to extract the hidden moods of users. For texts, they performed learning-based analysis by considering the forms and structures of a sentence as well as the words related to the human moods. In this work, a mood lexicon was built for text and emoticon analysis based on two well-known dictionaries: VSO dictionary and SentiStrength1 dictionary. The sentiments from images belong to a tweet were analyzed by SVM-based learning. Then moods of the respective sentence were aggregated per a tweet, furthermore per a day. Finally, the transition of user s daily moods was monitored for a long period and discriminate the users with depressive moods from others. They evaluated their approach with 3780 reviews and 2822 tweets. It was reported that the obtained accuracy perform better than the accuracy of SentiStrength with 4.4% to 28.0% improvements. Additionally, the experiment evaluated with 45 users proved the effectiveness of this method in finding depressive users. For text based sentiment analysis, two methods were used and compared. The first one was by using NB and the second by using SentiStrength API. SentiStrength API is a publicly available and was developed to compute the sentiment score for texts. It was reported that it performed better than using NB in this study. The performance of the visual sentiment analysis is better than the textual sentiment analysis such that an accuracy rate of 70% was reported using the visual sentiment analysis while 43% was also reported using the textual sentiment analysis. Combining text content (Sen-

tiStrength)with visual contents (SentiBank) resulted in improving the results such that an accuracy rate of 72% was obtained using Logistic regression.

Cai and Xia [113] utilized CNN to develop a multimodal sentiment analysis framework for textual and visual tweets. The architecture of their framework is composed of three CNN architectures: text CNN, image CNN and multi CNN. The multi CNN is fed as input by joint text-level and image-level representation. The overall architecture is composed three CNNs: one for texts, one for images and one which combined both of them to exploit the internal relation between text and image. The multi CNN took as input the joint text-level and image-level representation. It is composed of four connected layers and a softmax layer and does not contain convolutional and max pooling layers. It was reported that the accuracy rates obtained using the multi CNN outperformed other methods for multimedia sentiment analysis. Additionally, the multimodal accuracy rates higher than the individual modes.

Recent a study was conducted by Alqarafi et al. [123] for Arabic sentiment analysis based on combining textual and visual modality. A dataset of 42 videos are collected. BoW features are used to represent textual modality while smile, frown, head nod, and head shake based features are used to represent visual modality. The two modalities are fused at feature level fusion while SVM classifier is used. They reported an accuracy rate of 76.09%.

2.3.4 A-V Sentiment Analysis

Wang and Guan [124] first proposed a systematic approach based on audiovisual information to recognize emotions. Prosodic, Mel-frequency Cepstral Coefficient (MFCC),

and formant frequency features were extracted as audio features. Gabor wavelet features was applied to represent the visual information. combined audio and visual modality to recognize six emotions They constructed a dataset of videos recorded in lab environment

Wang et al. [119] introduced kernel cross-modal factor analysis approach to identify the optimal transformations for representing the coupled patterns between two different subsets of features. This approach was applied to analyze the cross-modal relationship between audio and visual features. The features were extracted from two emotion datasets: RML [118] and eNTERFACE [125]. A hidden Markov model (HMM) was applied as a classification method to detect emotions and to measure statistical dependence across the successive time segments. It was reported that the proposed method outperformed the simple feature and score fusion levels and performed better than the original features.

Recently, a study is presented by [126] to combine audio and visual modalities in video level analysis. Twenty one videos downloaded from YouTube and expressed in Arabic were prepared as a dataset. Voice energy, voice power, intensity and pitch were extracted as acoustic features and two main visual features based on smile and eye with considering only feature fusion level. Different classifiers were applied including: decision tree, k - nearest neighbor (K -NN), naive Bayes, SVM, and ANN. The highest results were reported using audio modality in all cases while combining audio and visual modalities causes dropping the results of audio modality in nearly all cases. This can be attributed for several reasons such as the small number of instances.

2.3.5 A-T-V Sentiment Analysis

To our knowledge, the first study addressed the three modality sentiment analysis was conducted by Morency and Mihalcea [94]. They found that it is a feasible task and can take advantage from the joint exploitation of audio, visual, and textual modalities. A dataset was built which is composed of 47 videos collected from YouTube. The videos cover different topics such as politics, electronics products reviews, religions, etc.

Wöllmer et al. [95] presented a multimodal framework for analysing sentiments in YouTube movie reviews. Institute for Creative Technologies Multi-Modal Movie Opinion (ICT-MMMO) dataset of 370 multimodal review videos was built in this study. In-domain and cross-domain data were considered and compared. For in-domain sentiment analysis, ICT-MMMO dataset was used. For cross-domain sentiment analysis, Metacritic dataset [127] were used for training and ICT-MMMO were used for testing. A set of 1,941 audio features were extracted and openSMILE tool was used for this purpose. Cyclic correlation-based feature subset selection (CFS) technique was used to reduce the number of features. Transcripts were generated automatically using Automatic Speech Recognition (ASR) system and also manually. A set of 1000 bag-of-words (BoW) and bag-of-n-gram (BoNG) textual features was extracted and then fed into linear SVM. Additionally, a set of 20 visual features was extracted and were reduced via feature selection technique to six features, in average. Audio and visual features were fusing to concatenate a feature vector at level feature and fed into bidirectional long short-term memory (BLSTM) recurrent neural networks. However, decision fusion level were considered for fusing text and audio-visual modalities. It

was reported that, the manual transcriptions more accurate than the generated automatically ones. It was also reported that the highest accuracy and F_1 of 73.2% was obtained using the combined textual and visual modality in case of in-domain settings which is followed by the textual modality such that an accuracy and F_1 of 73% were reported.

Pérez-Rosas et al. [96] presented a multimodal sentiment classification method to determine the sentiment expressed in utterance-level visual data-streams. Bag-of-words were used as textual features and prosody, energy, voicing probabilities, spectrum, and cepstral features were extracted using OpenEAR tool as acoustic features. While facial expressions were extracted as visual features and CERT [128] tool was used for this purpose. A Multimodal Opinion Utterances Dataset (MOUD) was created, in this work for Spanish product opinions. Feature level fusing method was considered. Analysis was performed in case of utterance level and video levels. It was reported that the combining different modalities performed better than the individual modalities in all cases while the video analysis level achieved higher results than the utterance level.

Rosas et al. [97] presented a multimodal sentiment analysis method for Spanish videos. They compiled a dataset of 105 videos and considered the visual, audio and textual modalities. BoW features were extracted for the text modality while pause duration, pitch, intensity, and loudness were extracted as acoustic features. Smile duration and Look-away duration series of features were extracted as visual features. These features were concatenated and fed into SVM. It was reported that the highest accuracy rate, 75%, was obtained using the combined features of the three modalities.

Another dataset of 37 videos were collected in this work for English reviews but more challenging. The videos don't only show the speaker but also the product they review. It was reported that textual modal sentiment analysis achieved the highest results with an accuracy rate of 64.94% which is higher than the multimodal approach, for the English dataset.

Poria et al. [98] presented a multimodal framework using visual, audio and text modalities to determine sentiment polarity of video clips. Feature-level and decision-level fusion methods were used to combine different modalities for MOUD dataset. The textual contents were translated to English using Google Translator. A deep CNN to extract features from text was proposed. The input to CNN was formed by constructing a 306-dimensional vector for each word. The vector was generated by concatenating word2vec dictionary of 300-dimensional vector and six parts of speech tags including: noun, verb, adjective, adverb, preposition, conjunction which extracted using Stanford Tagger. Each clip was split into frames and 68 facial characteristic points (FCPs) were extracted. Each facial expression was characterized by $68 \times 67/2 = 2,278$ distances. Six further face position coordinates were extracted for each frame to come up with 2,284 values per frame. A set of 4,568 visual features was obtained by calculating mean value and standard deviation over all frames of the clip. Additionally, 6,373 audio features were extracted using openSMILE software. A Multiple Kernel Learning (MKL) algorithm was applied to classify the multimodal combined feature vectors. It was reported that without translating into English an accuracy rate 68.56% was obtained for text modality. The best performance was obtained using the multimodal approach with applying feature selection techniques. The highest accuracy

rate of 88.60% was reported in case of feature level fusion which was higher than what was reported in [96].

Poria et al. [99] presented a multimodal sentiment analysis framework to combine text, audio and visual modalities. Both feature level and decision level fusion were applied and compared. The framework was evaluated using the YouTube dataset [94]. Several machine learning classifiers were used: NB, SVM, extreme learning machine (ELM) and artificial neural networks (ANN). It was reported that, the best performance was obtained using feature fusion level of three modalities, such that a precision of 0.782 and a recall of 0.771 were obtained.

Pereira et al. [105] combined audio, textual and visual information to determine tension level of news videos. As for visual features, they extracted visual intensity, participants field size and the prosodic features. Voicing probability, loudness and fundamental frequency were extracted as acoustic features while they used sentiment scores extracted from the closed caption as textual features.

Another approach for combining visual, acoustic and text modalities was presented by Poria et al. [106]. They proposed a temporal CNN for visual modality such that each pair of images at time t and $t + 1$ were combined into a single image. This was followed by inserting recurrent neurons hidden layers in the deep CNN model and then initializing the distributed time-delay weight matrix of recurrent neural networks (RNN) with the covariance of CNN output. OpenSMILE toolkit was adopted to extract pitch and voice intensity represented as 6,373 acoustic features. As for textual features, CNN was applied as trainable feature extractor similar to [98]. The most significant features were kept by applying features selection techniques.

Then, the features of each modality were concatenated (feature-level fusion) and fed to a MKL classifier. The method was validated on multimodal sentiment analysis dataset, MOUD and multimodal emotion recognition dataset, USC IEMOCAP. For sentiment classification (polarity detection) an accuracy rate of 96.55% was obtained using the multimodal approach which is higher than [96]. With considering four emotions (angry, happy, sad and neutral) from USC IEMOCAP dataset, an average accuracy rate of 76.85% was reported which is higher than the results reported in [103].

A method based on tensor fusion network has been, recently, introduced to learn modeling intra-modality and inter-modality dynamics [102]. It was tailored for the volatile nature of spoken language in online videos as well as accompanying gestures and voice. It was reported that the tensor fusion network method outperforms state-of-the-art approaches for both multimodal and unimodal sentiment analysis.

Poria et al. [100] utilized LSTM to enable utterances to capture contextual information from their surroundings in the same video. The method is evaluated on MOSI, MOUD, and IEMOCAP datasets. it was concluded that, the presented model results in improving the classification process with reporting 5-10% improvement in results over the state of the art.

Interested readers can be referred to recent and comprehensive literature reviews conducted in [13, 129].

2.3.6 Multimodal Sentiment Analysis Datasets

Due to the challenging settings of building datasets of multimodal sentiment analysis, few datasets were constructed and publicly available. In this section we define the

process of constructing multimodal sentiment analysis dataset with hopping this helps researchers in this area. These criteria includes the following:

- Diversity: it includes several aspects including the diversity in:
 - Domain: the content of the dataset belong to a certain domain (domain oriented) or general covers several domains such as politics, product reviews, religion views, etc. This criterion is determined based on the study and the application.
 - Gender: the content of the dataset are expressed by males and females or just expressed and described by a certain sender. Diversity in gender is preferable [94, 111].
 - Age: the topics in the contents of the dataset are expressed by people in different ages or not. Diversity in gender is recommended and add more challenges.
 - Dialect: some languages have several accents, for example Arabic has several dialects: Egyptian, Iraqi, Levantine, Gulf, Maghrebi, Yemeni. So, should the different dialects be considered or just the Standard language or focus on a certain dialect? Considering the Standard languages might be less challenging for processing since several resources are available. However, including several dialects might be more challenging or sentiment analysis. This criterion is determined based on the study and the application.
- Language: Each language has its phonemes, morphemes, lexemes, syntax, context, grammar and semantic. Some languages might have similar components

while others totally different. As shown in Table 2.4 English, Spanish, Chinese and Indian languages are considered. However, so far there is no Arabic multimodal sentiment analysis dataset.

- **Environment:** The data might be recorded in a lab environment or might be recorded in a real environment. The robustness and the effectiveness of the sentiment analysis systems rely on the recording environments. They need to be able to deal with the real-world variability and noises present in most video recordings [94, 111]. Most of the works collected their data from social media which were expressed in real-world environment. However, some domains might require preparing a special environment for recording the data. Some studies prepared their environments to record their datasets includes [116, 118, 130].
- **Polarity:** it determines the number of classes or categories and it relies on the task and the application. The more number of classes, the more challenging is. For example, the polarity of sentiment classification includes:
 - 2-class: Positive, Negative
 - 3-class: Positive, Negative, Neutral
 - 5-class: Strongly positive, Positive, Neutral, Negative, Strongly negative.
 - 8-class: Strongly positive, Positive, Weakly positive, Neutral, Weakly negative, Negative, Strongly negative, Uncertain.

On the other hand, an example of the polarity in the emotion recognition includes:

- All or some of: Happy, Angry, Sad, Surprised, Excited, Frustration, Disgust, Fear, Neutral
 - Tension: Low, High
- Transcription: by which the spoken words are extracted and other meta information such as the start time of each spoken utterance. Transcription can be conducted either manually, automatically or both manually and automatically. For some languages there are some tools to perform it automatically such as ASR and Kaldi [131] tools. Transcription in manual manner might be a tedious task and time consuming, yet it is reliable. This was confirmed by the study of [95] such that they evaluated and compared the textual features extracted from the manual and automated transcriptions. It was reported that the features extracted from manually generated transcriptions are more accurate and discriminate than the generated automatically ones. So, most studies considered manual transcription while the studies of [95, 103, 115] considered the automatically generated transcriptions.
 - Segmentation: to segment videos to frames and spoken utterances such that each utterance represent a certain class. Some studies considered the whole videos to express a certain sentiment or emotion while others segmented it to smaller levels such as utterance level. There are some criteria that be considered to utterance segmentation such as the long pauses [94]. This also can be conducted easily automatically using tools like Praat and OpenEAR [132]. This also might be related to the methodology of a study i.e, the analysis level of the study. Table

2.5 illustrates the distribution of the different levels among the corresponding studies.

- Annotation: by which each sample of the datasets is annotated or labeled to the corresponding class. The corresponding class depends on the considered polarity. Annotation are conducted either manually (by using professionals), automatically or weakly (using a tools with set of keywords, emoticons, etc.) or by using both methods. For example, a list of positive and negative sentiment keywords were used in [110] to query Getty Images. The returned images were labeled based on the sentiment labels of the keywords. Amazon Mechanical Turk (AMT) is a tool used to manually annotate samples in [110, 114]. VADER[122] was applied to weakly annotate tweets in [110]. Annotation might be based on the whole video or on the utterance after segmentation. Manually annotated is time consuming and tedious task but it is more reliable.
- Balance: balance here means is that the numbers of samples are distributed equally on the different classes. Nearly all datasets are not considered this aspect. Investigating the effect of balancing samples has not be considered in the reviewed studies.
- Other factors should be considered when constructing a multimodal dataset including:
 - The size of the dataset
 - The duration of video and audio

- The bandwidth of video and audio

Table 2.4 shows a summary of the adopted datasets on multimodal sentiment analysis and their categories based on important criteria. To our knowledge, so far, we are not aware of a work conducted on multimodal SAOM for Arabic language.

2.4 Discussions

With the rapid growth volume of resources on the Web, analysing this material becomes a crucial challenge. Much effort has been conducted to analysis the textual contents of these resources. These approaches showed their ability for opinion mining and analysis in different domains. However, these approaches are limited since they focus on a partial information source and ignore other resources of information. Since 2011, researchers started considering the other media sources including visual and audio contents and studying their effects when combined with the textual information. However, much efforts are required.

Multimodal sentiment analysis includes:

- Tri-modality systems: Audio, visual and textual.
- Bi-modality: Audio and visual, audio and textual as well as visual and textual.

Several methods are applied to combine these modalities in different levels. Feature, score, decision and hybrid fusion levels were applied. Selecting the effective fusion method relies on the data characteristics, the used methodologies and the application problem requirements. Most of the reviewed studies considered feature level

Table 2.4: Multimodal Datasets

Name	Ref	Modality	Genre	Used in	Labeled	Method	Language	Total	Polarity	Balanced	Evaluated
YouTube	[94]	Audio, Visual, Textual	Youtube Videos	General	[94, 99, 106]	Manual	English	47 videos, 498 utterances	3-class +, -, =	No	Yes
ICT-MMMO	[95]	Audio-visual	YouTube, ExpoTV	Review videos	[95, 106]	Manual	English	370 videos	2-class +, -	No	Yes
MOUD	[96]	Audio, Visual, Textual	YouTube videos	Product opinions	[96, 98, 106]	Manual	Spanish	80 Video, 498 utterances	3-class	No	Yes
-	[97]	Audio, Visual, Textual	Youtube, ExpoTV	General	[97]	Manual	Spanish	105 videos 550 utterances, 10k words	3-class	No	Yes
USC-IEMOCAP	[133]	Audio, Visual, Textual	dyadic interactions Video		[103, 106, 133]	Manual	English	12 hours of video	10-class	No	Yes
eNTERFACE	[125]	Audio, Visual, Textual	Videos	Reactions after listening stories	[104, 119, 120, 125]	Manual	English		7-class emotions		Yes
-	[105]	Audio, Visual, Textual	TV news-casts	News	[105]	Manual	English	520 videos	2-class Low tension and High tension	-	Yes
SentiBank Twitter	[114]	Visual, Textual	Tweets	General	[108, 113, 114]	Manual	English	470 +ve, 133 -ve	2-class	No	Yes
	[107]	Visual, Textual	Micoblog Sina-Weibo	General	[107]	Manual	Chinese	6000	2-class, 3-class		Yes
Sina Weibo	[134]	Visual, Textual	Micoblog Sina-Weibo	General	[109, 134]		Chinese	6171 messages (5859 with images)	3-class	No	Yes
Getty	[110]	Visual, Textual	Getty Images	General	[110]	Weakly	English	588,221 images	2-class	No	Yes
-	[110]	Visual, Textual	Twitter	General	[110]	Weakly using VADER	English	220,000 tweets	2-class	No	Yes
MOSI	[111]	Audio, Visual, Textual	YouTube videos	General	[111]	Manual	English	93 videos from 89 distinct speakers,	2-class Subjective, Objective, 10-class		
Flicker	[114]	Visual, Textual	Flicker	General	[114]	Weakly	English				
-	[135]	Visual, Textual	tweets	General	[113]	Manual	English	769 +ve, 500 -ve	2-class	No	Yes
-	[116]	Audio, Textual	Dialogue collected in lab environment	Daily lives Interactions with Game	[116]	Manual	Chinese	2,033 sentences	4-class	-	Yes
Telugu	[117]	Audio, Textual	YouTube	Telugu movie songs	[117]	Manual	Indian	300 songs	Happy and Sad	-	Yes

Table 2.5: Multimodal sentiment analysis level

Ref	Task	Mode			Analysis Level
		Visual	Audio	Textual	
[96]	PD	✓	✓	✓	Utterance, Video
[106]	PD, ER	✓	✓	✓	Utterance
[94, 98, 99, 111]	PD	✓	✓	✓	Utterance
[95]	PD	✓	✓	✓	Utterance (for audio, visual) and Video (for text)
[103, 104]	ER	✓	✓	✓	Utterance
[97]	PD	✓	✓	✓	Video
[105]	ER	✓	✓	✓	Video
[116]	ER	×	✓	✓	Utterance
[117]	ER	×	✓	✓	Three levels: Whole song, Beginning of song, Ending of song

fusion. Fusion at feature level is more complex, yet it is more affective for sentiment analysis. This might be due to its flexibility since it provide more information. Feature level fusion outperformed the decision level fusion level when compared in the studies of [98, 99, 120]. However, in two cases out of three presented in [110] the decision fusion level performed better than feature level fusion. Score level fusion is characterized by the simplicity in implementation and scalability [120]. However, it has not been obtained more attention for sentiment analysis compared with other levels.

It is shown that multimodality proves its feasibility and can work together to improve sentiment analysis and opinion mining. In nearly all studies, combining different modalities improves the results over the individual modalities and in most of them significant improvements were reported. The effects of fusing multimodalities are analyzed and classified into positive effect, negative effect and no effect. Table 2.7 illustrates the experimental results and the effects in the studies with considering the

best performance for those studies that were affected positively by applying the multimodality. It details the different results when there is no improvements. In some experiments of the studies of [95, 96, 110] combining different modalities did not produce improved results. Two cases out of five in [95], the combination of visual and textual modalities produced less accuracy rates than the textual modalities. Additionally, in two cases out of five there are negative effects when combining audio and textual modalities and no effect in one case. Besides, combining the three modalities has negative effects in four cases out of five. This is in addition to audio-visual modality produced higher results than the tri-modality. This might be referred to the heterogeneity of the proposed combination method in this study; the audio and visual modalities combined in feature level while the audio-visual modality and textual modality were combined using late level fusion. Additionally, audio-visual modality considered spoken utterance analysis level while the textual modality considered the movie review video. In [96] the textual modality performs better than combined audio and textual modality but the combination of three modalities resulted in improving the results over the single modalities. In [110], the accuracy rates of textual modality higher than the combined visual-textual modality in two cases out of six. Additionally, the precision rates of textual modality higher than the combined visual-textual modality in the six cases. Both early and late level fusion in the study of [110] improved results in terms recall and F1 in some cases. The superiority of textual modality in this study compared with visual modality, early and late fusion might be referred to the nature of annotation methods followed in this study. Weakly labeled method either using VADER or the list of keywords are text-based nature sentiment analysis.

So, they might be biased for the textual sentiment modality. This is confirmed for VADER in [110]. Another evidence for being these methods biased for textual sentiment analysis is that when samples were labeled using AMT workers in [110] the visual sentiment analysis is improved.

Sentiment classification task including polarity determination and sentiment intensity as well as emotion recognition have been only considered. Most studies were conducted on sentiment classification [94–99, 107, 109–113, 115]. Some of them focused on emotion recognition [103–105, 116–120]. The studies of [106, 114] addressed both sentiment classification and emotion recognition. Some studies built resources (e.g, datasets, lexicons) to conduct and evaluate their methodologies such as [94, 95, 114, 118].

Sentiment analysis subtasks such as building resources (datasets, lexicons, tools) for multimodal sentiment analysis still require more attentions. Transcription is an important task for building dataset especial textual information. It was performed manually or automatically. Performing the transcription manually is a tedious task and time consuming. However, it is reliable. Some tools are available to conduct the transcription automatically such as ASR. However, such a tool is inefficient when applied for sentiment analysis which is confirmed by the study of [95]. So, sentiment analysis is sensitive significantly to the errors generated by ASR when compared with the manual transcription [95]. Another subtask of sentiment analysis is the annotation of the contents. It is performed manually in most studies. It requires professionals to ensure the validity. Besides, it costs huge human efforts especially with large datasets. Annotation is also conducted automatically, also known as weakly labeled,

in the studies of [110] which is biased for textual modality as discussed. An attempt was conducted to alleviate the noisy nature of the weakly labeled method of the large-scale training images by You et al. [136] by presented a CNN-based architecture called PCNN (Progressively CNN).

There are two methods to develop multimodal sentiment analysis, namely machine learning-based systems and hypergraph-based system. Nearly all studies are applied machine learning techniques including single classifiers, ensemble classifiers and deep learning (see Table 2.6). SVM is the most applied classifier [95–97, 99, 104, 107, 111, 112, 114–117]. Other classifiers were applied including Logistic regression [107, 108, 112, 114] , HMM [94, 119], BLSTM [95], NB [99, 107, 112, 114, 117, 118], ELM [99], ANN [99, 104, 116, 118]. k-NN [104], GMM [116–118] and FLDA [118]. Although, SVM is the most used classifier and chives higher results, Logistic regression outperformed it in [108, 114].

The ensemble of classification algorithms is proposed to improve the produced results over the single classifiers. However, they should be combined in prober way [137]. Ensemble SVM was presented in [103]. An ensemble classifier of GMM, SVM, and MLP was applied in [116] based on a meta Decision Tree. It improves the base-classifiers' results. In [117], an ensemble classifier of SVM and NB and an ensemble classifier of SVM and GMM were presented and compared.

Deep learning is employed in several studies. CNN was applied in [98, 106, 109, 110, 112]. Since CNN uses the back propagation technique, it might get stuck in a local optimum. This issue was addressed in [98] such that CNN is employed with SVM instead of the back propagation technique. Applying CNN for textual modularity has

the advantage of that it combines CNN sentence model uses convolution as an operator to combine semantically-related word vectors and the convolution layers extract features in a hierarchical manner [106]. RNN was also applied in [106].

Hypergraph is another method to conduct multimodal sentiment analysis and it is just applied in [107]. It outperformed NB, SVM and Logistic regression in this study.

Feature selection is the process by which the most significant features are selected. It is a crucial process for classification problems. It is more important for multimodal recognition and analysis systems to eliminate redundant and noisy features specially when applying feature level fusion. Feature selection techniques also show their capability to handle the curse of dimensionality problem which might arise in multimodal systems. was applied in [95, 98, 106], Principal Component Analysis (PCA) in [98, 106, 118] and Stepwise method based on Mahalanobis distance was applied in [118]. A stepwise method based on Mahalanobis distance and PCA were applied and compared in [118]. The selected features using the stepwise method are more discriminant and less than those selected by PCA in that study.

2.5 Summary

- The research on the sentiment analysis of social media content is remarkably growing for constructing resources and investigating new ideas and techniques to address various challenges. Text based sentiment analysis approaches have proven to be extremely useful in the field of sentiment analysis. However, they suffer from the problem of domain, topic, temporal independent, etc. Therefore,

Table 2.6: Algorithms

Technique	Reference
SVM	[95–97, 99, 104, 107, 111, 112, 114–117]
Logistic regression	[107, 108, 114]
ANN	[99, 104, 116, 118]
NB	[99, 107, 114, 117, 118]
GMM	[116–118]
HMM	[94, 119]
BLSTM	[95]
ELM	[99]
k-NN	[104]
Ensemble	[103, 116, 117]
FLDA	[118]
Deep Learning	[98, 106, 109, 110]
Hypergraph	[107]

there is a need to address such issues through incorporating different sources and modalities.

- There is a need to perform a systematic study to evaluate different methods to address imbalance class problem.
- There are limitations to applying deep learning techniques to sentiment analysis and these gaps need to be filled. For more specific, we are not aware of an study that applies word embedding with LSTM and with combined CNN with LSTM. In addition, we are not aware of study that combines CBOW with CNN for Arabic sentiment analysis. This motivates us to present a method for Arabic polarity detection using CNN and LSTM along-side with sing word2vec based features (skip-gram and CBOW).
- There is not multimodal sentiment analysis dataset for Arabic. This requires us to provide a multimodal dataset for Arabic sentiment analysis in systematic

Table 2.7: Results and Affects “↑” means positive effect, “↓” means negative effect and “=” means no effect

Work	Modality			Settings	Results and Affects						
	Visual	Audio	Text		Visual	Audio	Text	V+A	V+T	A+T	V+A+T
[94]	✓	✓	✓		F1:0.439, P:0.449	F1:0.419, P:0.408	F1:0.43 P:0.431	-	-	-	F1:0.553, P: 0.543↑
[95]	✓	✓	✓	Manual&ICT- MMMO	61.2	64.4	73.0	66.2 ↑	73.2 ↑	72.3 ↓	72.0 ↓
[95]	✓	✓	✓	ASR&ICT- MMMO	61.2	64.4	63.7	66.2 ↑	61.5 ↓	65.0 ↑	62.1↓
[95]	✓	✓	✓	Cross: Manual &ICT-MMMO	61.2	64.4	71.2	66.2 ↑	71.1 ↓	71.1 ↓	70.9 ↓
[95]	✓	✓	✓	Cross: ASR &ICT-MMMO	61.2	64.4	61.0	66.2 ↑	63.0 ↑	64.4 =	63.9 ↓
[95]	✓	✓	✓	Open: Manual &ICT-MMMO	61.2	64.4	59.6	66.2 ↑	64.7 ↑	64.7 ↑	65.0 ↑
[96]	✓	✓	✓		50.66	53.33	73.33	61.33↑	74.66↑	72 ↓	74.66 ↑
[97]	✓	✓	✓		61.04	46.75	64.94	77.23↑	73.68↑	-	75↑
[98]	✓	✓	✓		76.38	74.22	79.77	83.69↑	85.46↑	84.12↑	88.60↑
[99]	✓	✓	✓		P:0.681 R:0.676	P:0.652 R:0.671	P:0.619 R:0.59	P:0.732 R:0.731 ↑	P:0.725 R:0.719 ↑	P:0.712 R:0.710 ↑	P:0.782 R:0.771 ↑
[103]	✓	✓	✓		51.25	60.9	48.55	-	-	67.43↑	69.35 ↑
[104]	✓	✓	✓		81.20	78.57	78.70	-	-	-	87.95↑
[106]	✓	✓	✓		94.50	74.22	79.77	95.68↑	96.21↑	84.12↑	96.55↑
[114]	✓	×	✓		-	n\ a	43	n\ a	72↑	n\ a	n\ a
[107]	✓	×	✓		-	n\ a	60.31	n\ a	86.77↑	n\ a	n\ a
[108]	✓	×	✓	CBOW-DA-LR	69	n\ a	75	n\ a	79↑	n\ a	n\ a
[109]	✓	×	✓	2class	76.3	n\ a	81.1	n\ a	82.6↑	n\ a	n\ a
[110]	✓	×	✓	Getty & Early fusion	A:0.732 P:0.747	n\ a	A:0.696 P:0.806	n\ a	A:0.763↑ P:0.778↓	n\ a	n\ a
[110]	✓	×	✓	Getty & Late fusion	A:0.732 P:0.747	n\ a	A:0.696 P:0.806	n\ a	A:0.769↑ P:0.785↓	n\ a	n\ a
[110]	✓	×	✓	Tweets& Early fusion	A:0.553 P:0.584	n\ a	A:0.722 P:0.746	n\ a	A:0.717↓ P:0.730↓	n\ a	n\ a
[110]	✓	×	✓	Tweets& Late fusion	A:0.553 P:0.584	n\ a	A:0.722 P:0.746	n\ a	A:0.604 ↓ P:0.634↓	n\ a	n\ a
[110]	✓	×	✓	Twitter Data, AMT&Early fusion	A:0.677 P:0.762	n\ a	A:0.688 P:0.832	n\ a	A:0.70↑ P:0.776↓	n\ a	n\ a
[110]	✓	×	✓	Twitter Data, AMT& Late fusion	A:0.677 P:0.762	n\ a	A:0.688 P:0.832	n\ a	A:0.716↑ P:0.799↓	n\ a	n\ a
[111]	✓	×	✓	Multimodal Dic- tionary	0.36	n\ a	0.46	n\ a	0.53 ↑	n\ a	n\ a
[113]	✓	×	✓	Multi CNN, DB1	0.773	n\ a	0.74	n\ a	0.78 ↑	n\ a	n\ a
[113]	✓	×	✓	Multi CNN, DB2	0.723	n\ a	0.77	n\ a	0.796 ↑	n\ a	n\ a
[115]	×	✓	✓	Cross-validation	n\ a	65.62	70.50	n\ a	83.33↑	n\ a	n\ a
[116]	×	✓	✓		n\ a	80.92	83.55	n\ a	85.79↑	n\ a	n\ a
[117]	×	✓	✓		n\ a	88.3	75.7	n\ a	n\ a	91.2↑	n\ a
[120]	✓	×	✓		-	n\ a	60.31	n\ a	86.77↑	n\ a	n\ a

manner.

- We are not aware of any study that addresses multimodal sentiment analysis for Arabic opinion videos. Consequently, this motivates us to conduct this research with developing the required resources such as the dataset.
- Sentiment analysis resources available to Arabic language are limited comparing to English.

CHAPTER 3

TEXTUAL ARABIC

SENTIMENT ANALYSIS

In this chapter, different types of textual features namely: *tf-idf*, LSA features as (hand-crafted features) and word embedding based features (as deep learning based features) are adopted and evaluated to detect polarities in Arabic microblogs. Several machine learning classifiers are used to evaluate the proposed features. Then, the class imbalance problem is addressed using different oversampling techniques.

This chapter also evaluates several deep learning methods based on convolutional neural networks and long short-term memory models for sentiment analysis of Arabic microblogs. Neural language models were trained using two different word2vec based techniques: CBOW and skip-gram. The top layer of those architectures are designed to include different approaches: static and non-static word initialization.

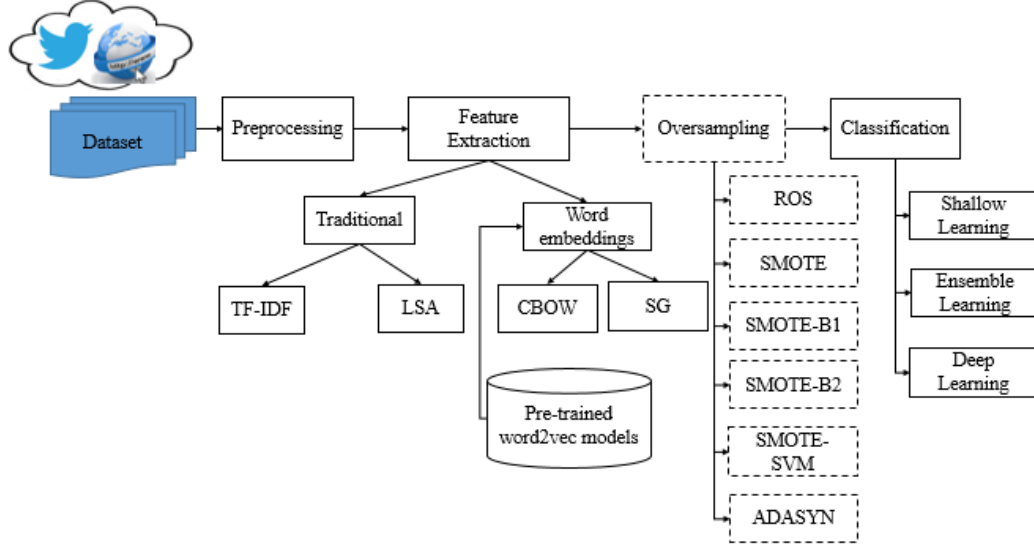


Figure 3.1: Layout of the textual Arabic SA approach

3.1 Framework of Textual Arabic SA

Figure 3.1 shows the layout of the followed methodology to detect sentiments using textual features. As a classification problem, the main steps are: data collection and preparation, text cleaning and preprocessing, feature extraction and classification. Other optional tasks, shown in dashed boxes in Figure 3.1, are considered including addressing class imbalance problem.

3.2 Feature Extraction

Feature engineering is a common and serious step for developing a machine-learning model in which each instance is mapped into a representation of its characteristics (features). Different textual features are considered and extracted including: *tf-idf*, LSA and two different forms of word embeddings.

3.2.1 Term Frequency-Inverse Document Frequency (*tf-idf*)

tf-idf is a popular term-weighting scheme to reflect how important a word to a document in a collection. A given text needs first to be represented as a matrix in which each row represent a unique word and each column represents a document (or tweet in our case) or other context. The count or frequency of how many times a term appears in a certain tweet is put in the corresponding cell, the value refers the Term Frequency (*tf*). Computing *tf-idf* relies on *tf* and Inverse Document Frequency (*idf*), such that $tf-idf = tf \times idf$. Given a set of N documents such that f_{ij} is the number of occurrences (frequency) of term i in document j , tf_{ij} can be computed by dividing f_{ij} by the number of terms in document j , (f_{*j}):

$$tf_{ij} = f_{ij} / f_{*j} \quad (3.1)$$

The *idf* is computed by taking the logarithm of dividing the total number of documents by the number of documents containing the term i , n_i ; thus it is computed from:

$$idf_i = \log_2 \frac{N}{n_i} + 1 \quad (3.2)$$

The *idf* gives a term that occurs in several documents less weight since it is assumed that this term is not discriminator as those occurring in few documents [138]. Finally, the *tf-idf* of term i in document j , is:

$$tf-idf_{ij} = (f_{ij} / f_{*j}) \times \log_2 \frac{N}{n_i} + 1 \quad (3.3)$$

The terms with the highest *tf-idf* score are often the informative and discriminating terms of the document topic. The computations are conducted while eliminating stop words. To avoid zero deviation, smoothing is applied to *idf* through adding one to the numerator and denominator

$$tf-idf_{ij} = (f_{ij}/f_{*j}) \times \log_2 \frac{1+N}{1+n_i} + 1 \quad (3.4)$$

Then Euclidean norm is applied to normalize the *tf-idf* vectors:

$$v_{norm} = \frac{v}{\|v\|^2} = \frac{v}{\sqrt{v_1^2 + v_2^2 + \dots + v_n^2}} \quad (3.5)$$

3.2.2 Latent Semantic Analysis (LSA)

LSA is a fully automated statistical approach of analyzing relations between terms and documents by the means of producing set of documents and their contained terms [139]. It is based on an unsupervised learning technique (clustering), and assumes that terms with common meaning occur in similar paragraphs. It builds a term-document matrix from a corpus, and aims at exposing some useful similarity structures for related text-analysis tasks and information retrieval.

tf-idf is an initial step of LSA. Then, Singular Value Decomposition (SVD) is applied to perform dimensionality reduction on the *tf-idf* vectors. The matrix generated in *tf-idf* is decomposed into the product of three other matrices [140]. The first matrix describes the original row entities as vectors of derived orthogonal factor values while the second matrix describes the original column entities in the same way.

The third matrix is a diagonal matrix containing scaling values such that when the three components are matrix-multiplied, the original matrix is reconstructed.

3.2.3 Word Embeddings

NLP applications that rely on hand-crafted features techniques suffer from several obstacles including the curse of dimensionality, being tedious and might be biased to features engineer problem due to high-dimensional features and requires high-computations. Embedding techniques are recognized as an efficient method for learning high-quality vector representations of words, terms or phrases from large amounts of unstructured text data. They refer to the process of mapping words, terms or phrases from the vocabulary to real-valued vectors such that elements with similar meaning to have a similar representation. There are different variations of embeddings tools including word2vec [66, 67], GloVe [76] and FastText. Our consideration in this study is word2vec.

Word2vec is a powerful tool developed by Google in 2013 [66, 67]. It efficiently computes word vector representations in high-dimensional vector space. Word vectors are located in the vector space where words that have similar semantic and share common contexts are mapped nearby each other in the space. In addition to syntactic information, similarity of word representations obtain semantic features such that semantic relationships are often preserved in vector operations on word vectors. For example, adding vector of (King) to vector of (Woman) and subtracting vector of (Man) is close to vector of (Queen). The word vector representations are proved to be efficient and successful technique in the applications of NLP such as text clustering

and classification and sentiment analysis.

Word2vec takes as input a large corpus of text and assigned a vector for each unique word in the corpus in the space. It has two neural network architectures: CBOW and Skip-Grams (SG) skip-grams. CBOW and SG have similar algorithms but the former is trained to predict a word given a context whereas the latter is trained to predict a context given a word. Figure 3.2 shows a shallow neural network model for CBOW and SG word embeddings. There are some parameters that need to be adjusted when training word2vec models such as sub-sampling, Dimensionality, and Context window. An important factor that affects the quality of word embedding is the dimensionality such that the higher the dimensionality, the higher the model quality until reaching some point. Higher dimensionality might result in more accurate models but more complex in terms of computational time and space. Another important parameter need to be taken into account when generating word embedding models is the size of the context window. It determines how many words before and after a given word would be included as context words of the given word. Determining the size of the context window depends on several criteria including the used technique (i.e., CBOW or Skip-grams), the genre of text used to learn word embedding models (tweets, paragraphs, articles, etc.).

Assume a sentence S of n words, $S = \{m_1, m_2, \dots, m_n\}$, where m_i is the i^{th} word in S . Let $x_i \in \mathbb{R}^d$ be the d -dimensional word vector corresponding to the i -th word in S . S is represented as $S = x_{1:n} = x_1 \oplus x_2 \oplus \dots \oplus x_n$. The sentence S is represented by a d -dimensional vector x_i , $x_i \in \mathbb{R}^d$ as an $d \times n$ matrix, where the element in the i^{th} row is corresponding to the word vector x_i . Each instance

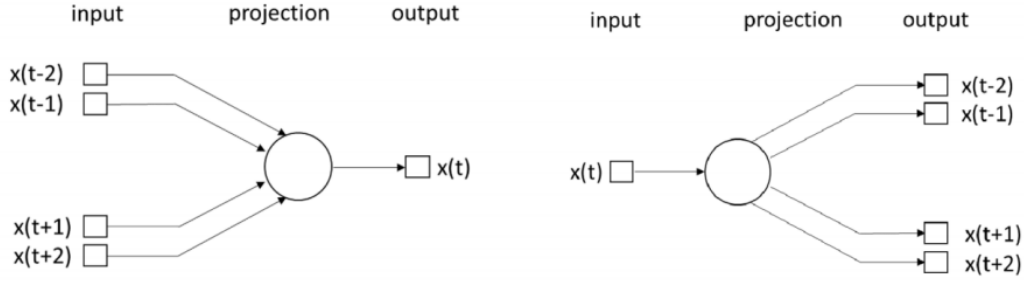


Figure 3.2: Neural network architectures for learning word2vec models (left: CBOW, right: Skip-gram)

or example in the dataset will be represented using a descriptor of dimensionality of $d \times n$.

In this study, the need is to represent the generated descriptor by a feature vector of dimensionality of $1 \times d$ and to be used as a textual based features to determine the polarity in Arabic texts. In other words, each generated $d \times n$ matrix need to be summarized as a feature vector to be fed into and learn a machine learning classifier (see Figure 3.3).

Since word embeddings based features of a sentences are represented as a matrix of $d \times n$, several ways can be utilized to summarize them or convert the word embeddings matrix into a textual feature vector for each sentence. Average (arithmetic mean) is a common way to summarize and generate a single feature vector from the word embeddings.

$$f_i = \frac{1}{n} \sum_{k=1}^n x_i, i = 1, 2, \dots, d \quad (3.6)$$

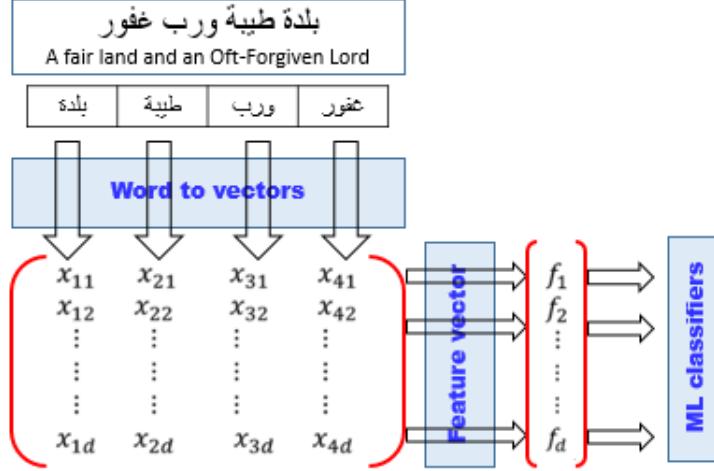


Figure 3.3: An example of sentence representation using word embeddings

3.3 Classification Methods

3.3.1 Single Classifiers

A) Support Vector Machine (SVM)

SVM [141] has been applied successfully in several NLP applications including sentiment analysis, authorship attribution, text classification, etc. SVM is a discriminative classifier represented, in its basic form, by a line separating a plane in two parts (hyperplane). The main objective is to minimizing the error and maximizing the margin hyperplane. Given training dataset of N points formed as: $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)$ alongside with (y_1, y_2, \dots, y_n) where \vec{x}_i is a p -dimensional feature vector and $y_i \in \{-1, +1\}$ is the corresponding label. Figure 3.4 represents features of binary class data separated using two hyperplanes: diagonal solid line and horizontal solid line. Mathematically,

the hyperplane is a set of points \vec{x} where:

$$\vec{w} * \vec{x} - b = 0 \quad (3.7)$$

where \vec{w} is the normal vector (perpendicular vector) to the hyperplane. The offset of the hyperplane from the origin along the \vec{w} is determined by:

$$\frac{b}{\|\vec{w}\|} \quad (3.8)$$

Points belong to positive class lay in the left of diagonal line or above the horizontal line:

$$\vec{w} * \vec{x} - b = 1 \quad (3.9)$$

While points belong to the positive class lay in the right of the diagonal line or below the horizontal line:

$$\vec{w} * \vec{x} - b = -1 \quad (3.10)$$

The distance between two hyperplanes is:

$$\frac{2}{\|\vec{w}\|} \quad (3.11)$$

The aim is at finding the maximum margin hyperplane to separate the two classes. To do so, \vec{w} should be minimized. In other words the distance between hyperplane and the closest point in class should be maximized. Figure 3.4 shows a separable data points and two hyperplanes. The optimal one is the diagonal because it has the

maximum margin. Another objective is that points should not fall into the margin. Towards this objective, there are constraints should be satisfied by each input vector \vec{x}_i :

$$\vec{w} \cdot \vec{x}_i - b \begin{cases} \geq 1 & y_i = 1 \\ \leq -1 & y_i = -1 \end{cases} \quad (3.12)$$

To be formulated as an optimization problem: "Minimize $\|\vec{w}\|$ s.t. $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$, for all $1 \leq i$

leqn. The classifier, $\vec{x} \mapsto \text{sgn}(\vec{w} \cdot \vec{x} - b)$. Therefore, the points lie closest to the hyperplane determine the max margin hyperplane and called support vectors.

Computing the SVM model for nonlinear separable data is an optimization problem to minimizing an expression of :

$$\frac{1}{n} \sum_{i=1}^n \zeta_i + \lambda \|\vec{w}\|^2 \quad (3.13)$$

s.t.

$$\vec{w} \cdot \vec{x}_i - b \begin{cases} y_i(w \cdot x_i - b) \geq 1 - \zeta_i & \forall i = 1, 2, \dots, n \\ \zeta_i \geq 0 & \forall i = 1, 2, \dots, n \end{cases}$$

where ζ_i is the smallest positive value that is satisfying $y_i(w \cdot x_i - b) \geq 1 - \zeta_i$,

$$\zeta_i = \max(0, 1 - y_i(w \cdot x_i - b)) \quad (3.14)$$

y_i is the i^{th} actual output and $w \cdot x_i - b$ is the predicted output. λ determines the trade-off between increasing the margin size and ensuring that the \vec{x}_i lies on the proper

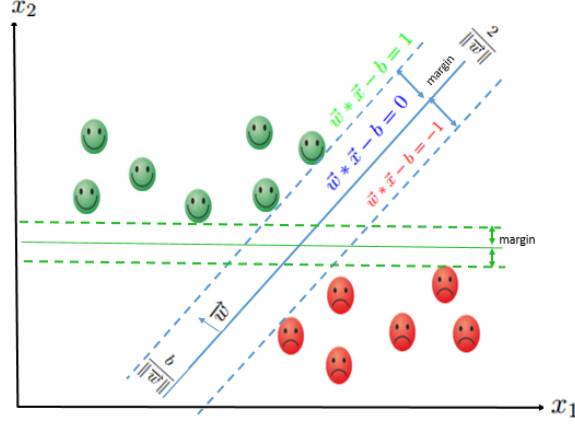


Figure 3.4: Hyperplanes and margins in SVM

side of the margin.

B) Logistic Regression (LR)

Logistic regression is a discriminative model aims at estimating the posterior probability $P(Y|X)$ from the training data [142]. Given a vector X of d attributes, $X = \{X_1, X_2, \dots, X_d\}$ with its label Y , and $Y \in \{0, 1\}$

$$P(Y = 1|X) = g(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}} \quad (3.15)$$

where

$$g(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

is the logistic function and

$$\theta^T X = \theta_0 + \sum_{i=1}^d \theta_i X_i$$

Since the sum of probabilities is equal to one,

$$P(Y = 0|X) = 1 - g(\theta^T X) = \frac{e^{-\theta^T X}}{1 + e^{-\theta^T X}} \quad (3.16)$$

By combining Equation 3.15 and Equation 3.16,

$$P(Y = yk|X; \theta) = (g(\theta^T X))^{yk} (1 - g(\theta^T X))^{1-yk} \quad (3.17)$$

where θ is the estimated parameter vector.

C) k -Nearest Neighbors (k -NN)

k -NN is a non-parametric method computes the distance between each training sample and the test sample. Several distance measurement methods can be used such as Euclidean Distance, Cityblock Distance, Minkowski Distance. The main parameters to be determined are the value of k and the type of distance measurement. The value of k is preferred to be odd such as $k = 1, 3, 5, ..n$. For two points a and b , with k -dimensions, the Euclidean Distance ED_{ab} can be calculated as:

$$ED_{ab} = \sum_{i=1}^k \sqrt{(a_i - b_i)^2} \quad (3.18)$$

where (x_1, y_1) and (x_2, y_2) are points in 2-dimensional space. The City-block Distance CD_{ab} can be calculated as:

$$CD_{ab} = \sum_{i=1}^k |a_i - b_i| \quad (3.19)$$

D) Naïve Bayes (NB)

NB classifier is a simple probabilistic conditional probability model that applies theorem of Bayes with strong (naïve) independence assumptions between the attributes. Give an instance of n features represented as $\vec{x} = (x_1, x_2, \dots, x_n)$ and the aim is to classify it as one of K classes C_k . NB assigns K probabilities to \vec{x} .

$$P(C_k | x_1, x_2, \dots, x_n) \quad (3.20)$$

$$P(C_k | x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (3.21)$$

The denominator is a constant and class independent and according to the naïve conditional probability, we can come up with:

$$p(C_k | x_i) = p(C_k) \prod_{i=1}^n p(x_i | C_k) \quad (3.22)$$

A class C_k with highest probability will be assigned as the predicted class for \vec{x} .

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}} p(C_k) \prod_{i=1}^n p(x_i | C_k) \quad (3.23)$$

There are different forms of NB based on the data distributions: Multinomial NB, Gaussian Naïve Bayes (GNB) and Bernoulli NB.

E) Decision Tree (DT)

A decision tree which is a rooted, directed tree similar to a flowchart, by which the input space is hierarchically divided until reaching a subspace associated with a class label [142]. Each internal node corresponds to a partitioning decision, and each leaf node is mapped to a class label prediction. Given training dataset \mathbf{X} of N points formed as: $(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n)$ alongside with (y_1, y_2, \dots, y_n) where \vec{x}_i is a p -dimensional feature vector and y_i is the corresponding label. The training set \mathbf{X} is split into a set of subsets $\mathbf{X}_s = \{X_1, X_2, \dots, X_k\}$ and $\bigcup_i X_i = \mathbf{X}$. Each parent node has a set of child nodes corresponds to a partitioning \mathbf{X}_s of the parent's data set, with the full data set associated with the root. X_i is composed $|X_{ij}|$ belong to class y_j . The probability that a randomly selected member of X_i belongs to class y_j is $p_{ij} = \frac{|X_{ji}|}{|X_i|}$.

Decision tree is recognized as an easy classification method to be implemented and understood. Various decision trees classifiers can be built from a set of given attributes easily. However, constructing an optimal decision tree is not straightforward [142]. It is computationally infeasible because of the exponential size of the search space.

Several algorithms based on a greedy strategy have been developed to build a reasonably accurate, albeit suboptimal, decision tree efficiently and in an acceptable time. The greedy strategy grows a decision tree by making a series of locally optimum decisions about which attribute to use for partitioning the data using some measurements such as Gini Index, Entropy (Information gain), and variance reduction. Decision tree classification method has several algorithms such as ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3), Classification and Regression Tree (CART) and SPRINT.

3.3.2 Ensemble Learning

A ensemble-based learning classifier consists on combining various learning algorithms to obtain higher accurate classifiers than single classifiers. Several approaches have been proposed to build ensemble classification methods, including *bagging*, *boosting*, *voting* and *stacking*. There are several criteria to characterize ensemble classifiers. When the base learners in the ensemble are identical then it is called homogeneous ensemble classifier. However, when the base learners are different it is called heterogeneous ensemble classifier.

A) Random Forest (RF)

Random forest classifier for an input vector X is composed of a set of tree predictors $h(X, \Theta_i), k = 1, 2, \dots$ of random vectors $\{\Theta_1, \Theta_2, \dots\}$; each random vector Θ_i is sampled independently and distributed equally [143]. The overall classification decision is made by taking the majority votes of the individual predictors.

B) Gradient Boosting (GB)

Gradient boosting generates additive regression models by sequentially fitting a base learner to current “pseudo”-residuals by least squares at each iteration [144, 145]. For a set of M weak base learners (BL), usually decision tree, the gradient boosting assigns a weight w_m for the m prediction of each base learner BL :

$$F(\mathbf{x}) = \sum_{m=0}^M w_m h(\mathbf{x}; p_m) \quad (3.24)$$

where $h(x; p_m)$ is the prediction function of a BL for an input vector \mathbf{x} with parameters p . w_m , and p_m are the resulting weight and parameters through training process for $\forall m = 1, 2, \dots, M$ such that:

$$(w_m, p_m) = \arg \min_{w, p} \sum_{i=1}^N \Psi(y_i, F_{m-1}(\mathbf{x}_i) + wh(\mathbf{x}_i, p)) \quad (3.25)$$

where $\Psi(y, F(\mathbf{x}))$ is the loss function to be minimized, and

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + wh(\mathbf{x}, p_m). \quad (3.26)$$

C) Voting-based Ensemble (VE)

Voting-based Ensemble (VE) is a meta-classifier for combining similar or conceptually different machine learning classifiers for classification via majority voting. The final class label might be made based on hard or soft voting. In case of hard voting, the final class is the most frequent classifier predicted by the classifiers. Given an input vector (x) and a voting based ensemble of m classifiers, the final class can be computed as:

$$\hat{y} = mode\{C_1(\mathbf{x}), C_2(\mathbf{x}), \dots, C_m(\mathbf{x})\} \quad (3.27)$$

For soft voting, the final class is predicted by calculating the average of the class-probabilities or class-scores.

$$\hat{y} = \arg \max_i \sum_{j=1}^m w_j p_{ij}, \quad (3.28)$$

D) Stacking-based Ensemble (SE)

Stacked generalization term refers to any method for feeding information from one set of generalization techniques to another before forming the final prediction [146]. Stacking based ensemble learning combines different base classifiers ($clf_1, clf_2, clf_3, \dots, clf_n$) as first-level classifiers using meta-classifier as the second-level classifier. Each base classifier is trained using the training set and the predictions of each base classifier ($p_1, p_2, p_3, \dots, p_n$) are fed into the second level as features to train the meta classifier. The final class is based on the meta-classifier prediction (P_f).

3.3.3 Deep Learning

A) Arabic sentiment analysis using CNN

A CNN architecture similar to Kim [68] with minor changes is investigated. Assume a sentence S of n words, $S = \{m_1, m_2, \dots, m_n\}$, where m_i is the i^{th} word in S and the task is to predict the sentiment polarity as *positive* or *negative*. The sentence S is represented by an $n \times k$ matrix, where the element in the i^{th} row corresponds to a k -dimensional vector $x_i \in \mathbb{R}^k$ of the i^{th} word. To conduct convolution operation, a filter $w \in \mathbb{R}^{h \times k}$ is applied to a window of h words to generate a new feature. For each possible window in the sentence $\{x_{1:h}, x_{2:h+1}, \dots, x_{n-h+1:n}\}$, the filter is applied to each possible window of words in the sentence to produce a feature map $c = [c_1, c_2, \dots, c_{n-h+1}]$, which $c \in \mathbb{R}^{n-h+1}$. The next layer is a polling operation such as max, average or L_2 -norm is applied to the feature map. Max polling is the most common one and takes the maximum of feature map, i.e. $\hat{c} = \max\{c\}$. Average

pooling was often used historically but has recently fallen out of favor compared to max pooling in computer vision, especially object recognition [147]. We validated this claim by conducting some experiments using Arabic Sentiment Tweets Dataset (ASTD) where max pooling performed better than the average pooling operation.

In order to generate multiple features, multiple filters are used with different window sizes. This forms a vector $z = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_m]$, where m is the number of filters, in the penultimate layer, which is then passed to a fully connected soft-max layer. The final output is the probability distribution over classes. Although deep neural networks are very powerful machine learning systems, a main problem related to them due to a large number of parameters is overfitting. Additionally, these networks are slow to use when they are large; making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. This problem is addressed by randomly dropping out a proportion p of the hidden units in the penultimate layer during training [148]. In forward propagation the output unit y without dropout is $y = w.z + b$, while with dropout it becomes $y = w.(z \circ r) + b$, such that $r \in \mathbb{R}^m$ is a vector of Bernoulli random variables with probability p of being 1, and \circ is element-wise multiplication operator. During testing, the learnt weight vectors are scaled by p such that $\hat{w} = pw$ then \hat{w} is used to score the testing sentences.

The main steps of the adopted CNN method is shown in Figure 3.5. Three convolutional filters (3, 5, 7) are adopted with using max-over-sampling pooling filter since it reflects the most significant feature [68]. The dropout rate is set to 0.5, and a sigmoid function is applied to generate the final classification.

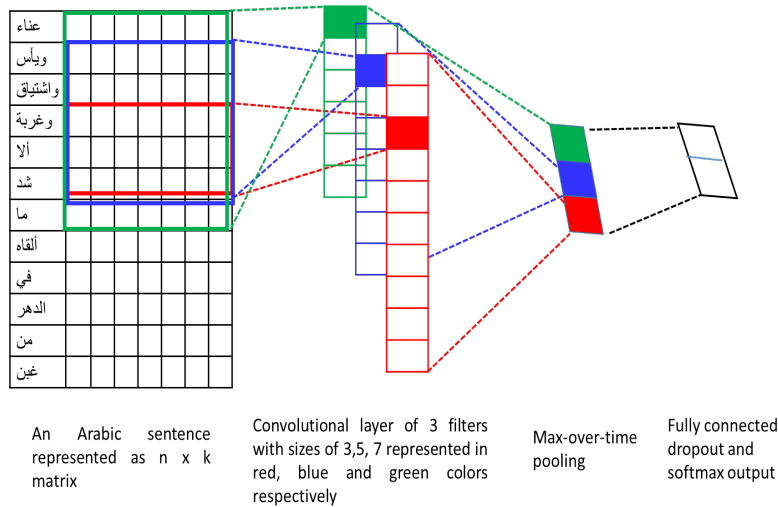


Figure 3.5: Adopted CNN architecture for Arabic sentiment analysis

B) Arabic sentiment analysis using LSTM

RNN is a popular type of artificial neural networks for sequences modeling. In their basic form, RNN can be considered as a densely connected neural network with feeding back the output of the hidden layer to itself. RNN has the ability to keep important part of the information to memory such as the words order relevant to each sentence. Therefore, RNN is recommended to model context dependencies in inputs of arbitrary length so as to create a proper composition of the input. RNN has been applied successfully in several NLP applications such as machine translation, speech recognition, image captioning, and language modeling, etc. In contrast to other supervised machine learning approaches, the order of observations has to be preserved when training RNN models. Sequence type of problems are classified into: sequence predictions (such as weather forecasting, stock market prediction and product recommendation), sequence classification (such as sentiment analysis, anomaly detection and weather

forecasting), sequence generation (such as text generation, music generation and image captioning) and sequence-to-sequence prediction (such as multi-step time series forecasting, text summarization and language translation).

RNN could not successfully train to keep to memory information for over long sequences because the gradients tend to either vanish or explode with serious impacts [149]. RNN has several forms including LSTM [71], Gated Recurrent Unit (GRU) [150], etc. Both LSTM and GRU are developed to struggle the vanishing gradient issue through gating mechanism. LSTM contains “memory cell” unit to keep information over larger input sequences. This cell is composed of four main components: input gate, a forget gate, a recurring cell state, and an output gate. GRU is introduced to make each recurrent unit to adaptively capture dependencies of different time scales. It has gating units that modulate the flow of information inside the unit, similar to LSTM. It differs from the latter in that it does not have a separate memory cells [151].

Four paradigms of LSTM recurrent neural network models are proposed to predict the sentiment polarity of Arabic text. Considering the opinion as a word sequence, LSTM has the advantage of recalling long-term spatial and temporal dependencies by linking past contexts to present one. The models considered here are as follows:

- Simple LSTM: Here, each word m_i is represented using one-hot encoding. LSTM model then takes this vector and converts it into a *word embedding* dependent vector.
- CNN-LSTM: LSTM layer is added to a CNN model (CNN-LSTM).

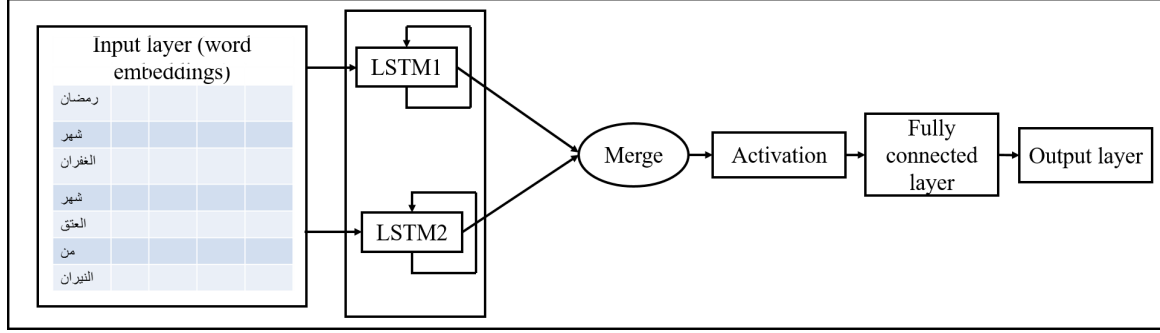


Figure 3.6: Combined LSTMs for Arabic sentiment analysis

- Stacked LSTM: Three LSTM layers are stacked on top of each other allowing the model to learn higher-level temporal representations. The first two LSTMs return their full output sequences, but the last one only returns the last step in its output sequence, thus dropping the temporal dimension (i.e. converting the input sequence into a single vector).
- Combined LSTM: an architecture based on LSTM is proposed by combining two LSTMs with dropout probabilities of 0.2 and 0.5, respectively. Different combination methods are investigated including: summation, multiplication and concatenation. The layout of this model is shown in Figure 3.6.

3.4 Handling Class Imbalance Problem

The imbalanced class problem has been addressed in several areas at the data level and/or algorithmic levels. Sampling techniques have been proposed to solve the imbalanced class problem at the data level to improve the predictive modeling capability. These techniques include oversampling, undersampling, and hybrid approaches. Over-

sampling techniques aim at balancing dataset through replicating or generating synthetic instances of the minority class. These technique vary in the way they generate synthetic instances. The focus of this chapter is on evaluating different oversampling techniques to address the imbalance class problem in sentiment analysis datasets:

- Random Oversampling (ROS): it is a non-heuristic method that over-samples the minority class in means of duplicating minority class examples, randomly or generating new examples from existing ones. Since this method makes exact copies of existing examples, this might lead to overfitting [152].
- SMOTE [153] over-samples the minority classes by adding synthetic samples based on feature-space similarities between existing minority examples. For each example x_i , where $x_i \in S_{min}$ (the minority class), it considers its k -nearest neighbors. SMOTE computes the difference between the sample under consideration and its nearest neighbor and multiplies it by a random number between zero and one. Figure 3.7 depicts a case of SMOTE with $k = 5$.
- Borderline-SMOTE [154]: SMOTE-based oversampling is performed on the borderline area since this area are very important to estimate the optimal decision boundary. A minority class sample is selected for oversampling in case of more than a half of it its m nearest neighbors belongs to the majority class. It has two variations Borderline-SMOTE1 and Borderline-SMOTE2. SMOTE-Borderline-1 (SMOTE-B1) generates synthetic samples from each sample in the borderline area and its positive nearest neighbors in the minority class. While, SMOTE-Borderline-2 (SMOTE-B2) considers the nearest negative neighbor in the ma-

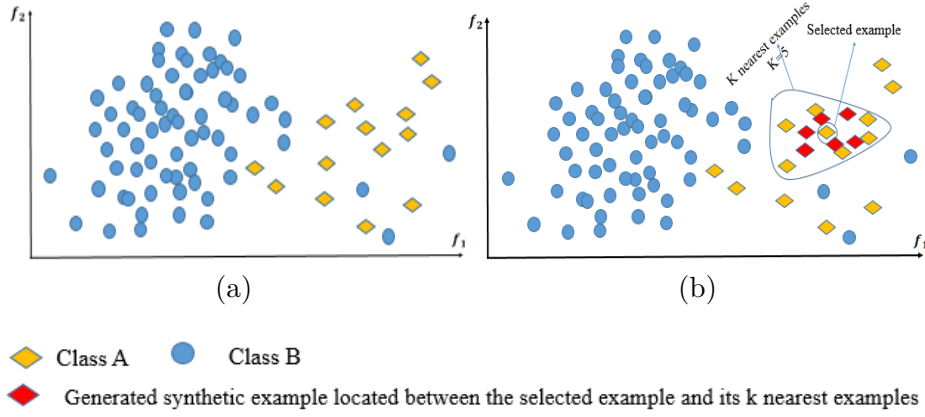


Figure 3.7: (a): Distribution of imbalanced dataset (b) Synthetic examples generated using SMOTE

jority class to generate synthetic samples.

- Support Vectors SMOTE [155] (SMOTE-SVM): it is another method that considers the borderline of the minority class where synthetic examples of minority class are generated along the decision boundary. The borderline area is approximated by the support vectors which are obtained through training a standard SVMs classifier on the original training set. Two methods are suggested to oversample a minority class instance in the support vector: the interpolation or extrapolation technique. With extrapolation, synthetic, samples are generated to expand minority class area toward the majority class. Applying the appropriate technique relies on the density of majority class samples around the minority class instance. The extrapolation technique is applied to oversample an instance of minority class in case of samples of the majority class around it are less than a half of its nearest neighbors and the new synthetic instance (X_{new}^+) is generated

as follows:

$$X_{new}^+ = sv_i^+ + p(sv_i^+ - nn[i, j]) \quad (3.29)$$

where sv_i^+ is a support vector, $nn[i, j]$ is the j^{th} positive nearest neighbor of sv_i^+ ; p is a random number in the range $[0, 1]$. Otherwise, the interpolation technique is applied similar to SMOTE to generate a synthetic instance, as follows:

$$X_{new}^+ = sv_i^+ + p(nn[i, j] - sv_i^+). \quad (3.30)$$

- ADASYN [156]: its idea is to generate synthetic examples for the examples belong to the minority class that are harder to learn without considering the easier ones. It uses a weighted distribution for different minority class examples according to their level of difficulty in learning.

3.5 Performance Evaluation

3.5.1 Evaluation methods

There are different methods to evaluate the proposed machine learning models, including:

1. Hold-out evaluation method: this method is recommended with large datasets.

The dataset is mostly divided into three independent subsets: training, testing and/ or validation datasets. The training set is used to build predictive models.

The validation set is a subset for evaluating the performance of model generated

in the previous phase. It serves as a test set to fine tuning model's parameters and selecting the best-performing model. The validation set might ignore in some situations depending on the problem, domain and applied algorithms. The third set is the testing set or unseen instances. The testing set is to evaluate the likely future performance of a model.

2. k -fold cross validation evaluation method the dataset is divided into k mutually exclusive and equal-sized subsets and the classifier is trained using $k - 1$ subset and test using the rest set. This process is repeated k runs with assuring all data is used for testing. With each run the error is calculated, the average error is calculated to evaluate the model. It is recommended with a small size of datasets.
3. Leave-one-out validation: it is a special case of cross validation method when k is equal to the size of samples. All data examples/ instances except one are used for training. The remaining instance is used for testing. This is repeated equal to the number of samples till testing all examples. It recommended with very small dataset and when the most accurate estimate of a classifier's error rate is required. It is more expensive computationally.

In this study, both hold-out evaluation mode and 10-fold cross validation are considered. Specifically, most of our experiments are evaluated using k -fold cross validation except in case of deep learning based approaches.

Table 3.1: Confusion matrix definition		
	Positive prediction	Negative Prediction
Positive class	TP: true positive	FN: false negative
Negative class	FP: false positive	TN: true negative

3.5.2 Evaluation metrics

In case of binary classification problem, the quality of each classifier is generally expressed in terms of the confusion matrix illustrated in Table 3.1 such that:

- True Positive (TP) indicates the number of positive examples that are classified correctly.
- False Positive (FP) indicates the number of negative examples that are classified incorrectly.
- True Negative (TN) indicates the number of negative examples that are classified correctly.
- False Negative (FN) indicates the number of positive examples that are classified incorrectly.

A variety of aggregate measures computed of these four quantities are very common in information retrieval, medical diagnosis and machine learning literature.

Accuracy is the most applied evaluation measure in the literature. However, it does not consider as a perfect measure, for imbalanced data-sets, since it does not differentiate between the numbers of examples classified correctly of different classes [152]. This probably results in incorrect conclusions. For example, assumes an IR of a dataset is 9.5 and all examples are classified incorrectly as negative class, then the accuracy will

be 95%.

$$Accuracy = \frac{\text{number of instances classified correctly}}{\text{total number of instances}} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.31)$$

$$Prc = \frac{TP}{TP + FP} \quad (3.32)$$

$$Rec = \frac{TP}{TP + FN} \quad (3.33)$$

F_1 is a popular measure used by the machine learning research community and is defined as:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.34)$$

GM is another good indicator for imbalanced dataset since it is independent of the examples distribution between classes. This measure tries to maximize the accuracy on each of the two classes while keeping their accuracies balanced [157, 158].

$$GM = \sqrt{TP_{rate} \cdot TN_{rate}} = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}} \quad (3.35)$$

Matthews Correlation Coefficient (MCC) is recommended as another perfect metric for imbalanced dataset. It is computed as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3.36)$$

All aforementioned evaluation measures range from 0 to 1, except MCC which ranges from -1 to 1. Receiver Operating Characteristic (ROC) Curve is used to combine

measures of positive class and negative class and visualize trade-off between benefits (TP_{rate}) and costs (FP_{rate}). TP_{rate} represents the x -axes and FP_{rate} represents the y -axes. The best classifier will score in the upper left corner, coordinate (0,1), of ROC space ($FP_{rate}=0, TP_{rate}=100\%$). However, the worst possible prediction method will score in the bottom right corner, coordinate (1,0), of ROC space ($FP_{rate}=100\%, TP_{rate}=0$). For balanced datasets, a random classifier would give a point somewhere along the diagonal line from the coordinate (0,0) to the coordinate (1,1) ($FP_{rate}=TP_{rate}$) since the model will throw up positive and negative examples at the same rate.

The AUC is computed by getting the area of the graphic to provide a single measure of a classifier's performance for evaluating which model is better on average [159]. An ideal classifier has an AUC of one.

$$AUC = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n V \quad (3.37)$$

such that i index counts the true positive samples (m) and j index counts the true negative samples (n). For each sample i, j the predicted probability (scores) is p_i, p_j respectively while V is:

$$V = \begin{cases} 1 & \text{if } p_i > p_j \\ 0 & \text{otherwise} \end{cases}$$

Table 3.2: Description of the datasets used in the experimental study

Dataset	Positive	Negative	Total
ASTD	665	1496	2161
GS-dataset	876	1941	2817

3.6 Experiments

3.6.1 Experimental Settings

A binary sentiment classification for two publicly available datasets collected from twitter are considered. The first dataset is ASTD [49] and the second dataset is Arabic Gold-Standard dataset [48] (GS-dataset) which was collected in 2014. Both datasets are described in Table 3.2

As for preprocessing step, the following operations are performed: removing non-Arabic symbols, diacritical marks, punctuation marks, removing duplicate character, and normalizing Alefs and Ta-Marbotah

tf-idf and LSA features are extracted from the aforementioned datasets to from two different feature vectors per each evaluated dataset. As for *tf-idf* and LSA, the stop words were eliminated where a set of Arabic stop-words list¹ was applied. The list contains all forms of stops words around 10390 forms. Figure 3.8 shows an example of an Arabic stop word with its different forms. For LSA features, SVD is used with 100 components. Therefore, 100 features are extracted for each instance per dataset.

Word embedding features computed based on pretrained CBOW and skip-grams models in [78]. For those language models, different experimental settings were investigated, including embedding dimensions of 100, 200 and 300 with different size of windows. The word2vec language model was learned using a large Arabic corpus

إن		
Prefix	Suffix	Affix
فإن، كان،	إنه، إنها، إنهم، إنهن، إنهما، إنك، إنكما، إنكم، إنكن،	فإنه، فإنها، فإنهم، فإنهن، فإنهما، فإنك، فإنكما، فإنكم، فإنكن، كانه، كانها، كانهم، كانهن، كانهما، كانك، كانكما، كانكم، كانكن،

Figure 3.8: Example of an Arabic stop-word with its different forms through prefixes, suffixes and affixes

of around 3.4 billion words and a vocabulary of 2.2 million words written in modern standard Arabic and dialectal Arabic. A feature vector is generated for each sample by averaging the embeddings of that sample. For each dataset four type of feature are extracted: *tf-idf*, LSA and two forms of word embeddings (CBOW and skip-grams). The size of *tf-idf* vector for ASTD dataset is 3674 features. While, the size of *tf-idf* vector for GS-dataset is 4063 features. The size of LSA vector for all considered datasets are 100 features. For word embedding base features the generated vectors are with size of 300 features for each dataset. *tf-idf* and LSA are consider as traditional or hand-crafted features. These two forms of features are widely applied in NLP tasks and report acceptable results.

Several base and ensemble classifiers are considered to generate machine learning models in this study. For base classifiers, SVM, GNB, *k*-NN, and DT. While, Random Forest (RF), Gradient Boosting (GB), VE, and Stacking-based Ensemble (SE) are used as ensemble classifiers. The classifiers are implemented in Python using scikit-

Table 3.3: Summary of evaluated base and ensemble classifiers

Type	Classifier	Description
Base	SGD	Regularized linear model with stochastic gradient descent learning, L1 norm regularization, and LR loss
	SVC	Linear C-SVM from LibLinear, C =1, and L2 penalty
	GNB	Gaussian Naïve Bayes
	NN	k-NN with k = 1
	DT	Decision tree with Gini index and min sample split = 2
Ensemble	RF	Random forest with 100 trees
	GB	Gradient boosting
	VE	Soft voting ensemble with base classifiers: SGD, Logistic Regression (LR) and LRCV
	SE	Stacking ensemble with base classifiers: SGD (LR loss, L1 norm regularization) and SGD(L2 norm regularization, Hinge loss), SVM

Table 3.4: Training parameters of Arabic word embeddings

Model	Dimensionality	Window	Sampling	Negative	Min_count	Iterations
CBOW	300	10	0.0001	10	5	15
Skip-Gram						

learn package [160]. Table 3.3 summarizes the adopted parameters for each classifier.

Furthermore, various deep learning models are evaluated using two datasets of Arabic tweets: the ASTD balanced dataset preprocessed by Dahou et al. [78] is used. Arabic sentiment analysis (ArTwitter) [50], which consists of 2000 Arabic tweets is also used. An Arabic corpus of around 190 million words compiled from various sources (Quran-text, Watan-2004, CNN-Arabic, BBC-Arabic and consumer review) [77] is used to train CBOW and SG. These models were generated using Gensim tool with the parameters described in Table 3.4.

For implementation of deep learning models, the Keras deep learning package with Theano backend is used. While Gensim package is utilized for implementing word embedding models. All experiments are implemented on Python. Each of the proposed deep learning architectures in case of CBOW or SG are experimented with two model variations for word initialization, following Kim [68]:

- Static word embedding initialization: All words are kept static with values of pre-trained vectors from word2vec while the other parameters of the model are learnt for each task.

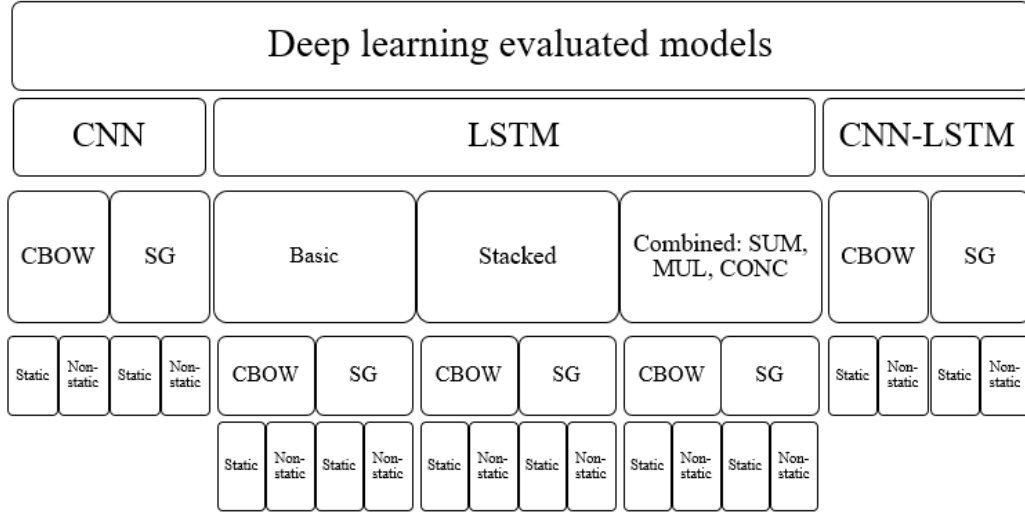


Figure 3.9: The evaluated deep learning models per each dataset.

Table 3.5: Performance of *tf-idf* and LSA features for ASTD dataset.

Clf	tf-idf				LSA			
	Acc	F_1	MCC	GM	Acc	F_1	MCC	GM
SGD	77.09 \pm 2.28	71.32 \pm 4.20	44.35 \pm 7.11	70.86 \pm 4.86	75.71 \pm 2.55	70.74 \pm 3.71	43.21 \pm 6.87	71.06 \pm 4.79
SVC	78.25 \pm 2.01	71.92 \pm 3.07	45.69 \pm 5.62	70.50 \pm 3.03	76.81 \pm 2.25	70.25 \pm 3.29	42.11 \pm 6.18	69.04 \pm 3.19
GNB	75.71 \pm 1.89	69.91 \pm 2.80	40.46 \pm 5.26	69.13 \pm 2.84	72.93 \pm 3.76	67.90 \pm 3.78	36.06 \pm 7.53	67.70 \pm 3.45
NN	51.63 \pm 4.42	49.27 \pm 3.64	2.30 \pm 6.51	51.23 \pm 3.48	68.77 \pm 2.31	64.63 \pm 2.38	29.89 \pm 4.96	65.42 \pm 2.75
DT	67.52 \pm 4.28	63.08 \pm 3.52	27.18 \pm 6.73	63.79 \pm 3.17	67.29 \pm 1.93	62.07 \pm 2.44	24.34 \pm 5.04	62.33 \pm 2.69
RF	76.26 \pm 2.23	70.12 \pm 2.91	41.41 \pm 5.82	69.12 \pm 2.85	75.20 \pm 2.03	63.89 \pm 3.93	35.27 \pm 6.21	63.13 \pm 3.27
GB	74.41 \pm 1.33	61.14 \pm 2.40	32.91 \pm 4.78	61.02 \pm 1.83	75.48 \pm 1.42	66.31 \pm 2.56	36.98 \pm 4.32	65.12 \pm 2.29
VE	78.02 \pm 2.01	71.21 \pm 3.65	44.92 \pm 6.06	69.86 \pm 3.69	77.32 \pm 2.12	71.35 \pm 3.86	44.16 \pm 6.71	70.52 \pm 4.35
SE	76.67 \pm 2.13	71.17 \pm 3.18	44.39 \pm 6.05	70.92 \pm 4.17	55.96 \pm 20.54	51.49 \pm 19.94	19.07 \pm 25.43	59.77 \pm 12.15

- Dynamic/ non-static word embedding initialization: All parameters are learnt and fine-tuned including words vectors for each task.

Consequently, 28 main models described in Figure 3.9 are generated for each dataset.

Hold-out evaluation method is considered to evaluate the word embedding models.

The datasets are divided as 75% for training and validation and 25% for testing.

3.6.2 Feature Techniques Evaluation

Table 3.5 presents the results of *tf-idf* and LSA features extracted from ASTD dataset using the aforementioned classifiers. *tf-idf* archives higher results than LSA in most

Table 3.6: Word embedding based features performance using ASTD dataset. The highest results are shown in bold font

Clf	CBOW				Skip-gram			
	Acc	F_1	MCC	GM	Acc	F_1	MCC	GM
SGD	83.99 \pm 1.97	81.14 \pm 2.37	63.01 \pm 4.38	81.32 \pm 3.16	82.70 \pm 2.74	79.78 \pm 2.85	60.73 \pm 4.85	80.25 \pm 3.25
SVC	84.59 \pm 1.35	80.78 \pm 1.85	62.60 \pm 3.35	79.26 \pm 2.00	84.54 \pm 1.89	80.67 \pm 2.77	62.33 \pm 4.92	79.14 \pm 2.92
GNB	77.79 \pm 1.93	74.10 \pm 2.00	48.34 \pm 3.93	74.27 \pm 1.94	76.59 \pm 1.45	72.21 \pm 1.76	44.54 \pm 3.52	72.02 \pm 1.89
NN	77.55 \pm 2.22	72.15 \pm 3.19	45.05 \pm 6.11	71.25 \pm 3.27	77.51 \pm 2.52	71.86 \pm 3.30	44.73 \pm 6.52	70.84 \pm 3.30
DT	73.54 \pm 3.37	69.12 \pm 3.93	38.33 \pm 7.84	69.28 \pm 4.03	70.76 \pm 3.24	65.69 \pm 4.21	31.49 \pm 8.38	65.81 \pm 4.31
RF	81.73 \pm 1.90	76.65 \pm 2.63	54.95 \pm 4.94	74.94 \pm 2.59	81.58 \pm 1.96	75.85 \pm 2.80	54.32 \pm 5.21	73.83 \pm 2.78
GB	83.66 \pm 1.56	79.76 \pm 1.90	60.34 \pm 3.75	78.38 \pm 1.88	83.02 \pm 2.04	78.72 \pm 2.69	58.47 \pm 5.19	77.21 \pm 2.73
VE	85.01 \pm 2.01	81.54 \pm 2.75	63.88 \pm 4.91	80.39 \pm 3.16	84.77 \pm 1.66	81.29 \pm 2.43	63.27 \pm 4.19	80.18 \pm 2.78
SE	84.22 \pm 1.24	80.53 \pm 2.14	61.95 \pm 3.66	79.42 \pm 2.94	83.71 \pm 2.46	80.24 \pm 2.94	61.40 \pm 5.21	79.54 \pm 3.18

Table 3.7: Traditional features performance using GS-dataset

Clf	tf-idf				LSA			
	Acc	F_1	MCC	GM	Acc	F_1	MCC	GM
SGD	78.17 \pm 2.62	71.59 \pm 3.53	46.59 \pm 6.23	70.26 \pm 3.19	74.91 \pm 5.38	68.80 \pm 3.88	41.97 \pm 6.46	68.58 \pm 2.96
SVC	79.73 \pm 1.63	74.02 \pm 2.27	49.94 \pm 4.32	72.39 \pm 2.21	78.06 \pm 2.78	71.23 \pm 3.60	45.29 \pm 7.34	69.61 \pm 3.27
GNB	75.43 \pm 2.17	70.56 \pm 2.72	41.38 \pm 5.36	70.12 \pm 2.74	70.18 \pm 3.32	66.18 \pm 3.43	32.65 \pm 6.81	66.71 \pm 3.41
NN	51.55 \pm 2.99	50.68 \pm 2.88	8.57 \pm 6.12	54.57 \pm 3.26	68.90 \pm 3.20	63.97 \pm 3.13	28.10 \pm 6.36	64.06 \pm 2.94
DT	66.95 \pm 3.15	64.47 \pm 3.01	30.91 \pm 5.72	66.46 \pm 2.99	67.41 \pm 1.43	62.59 \pm 1.70	25.36 \pm 3.51	62.88 \pm 1.91
RF	76.71 \pm 1.87	72.31 \pm 2.38	44.79 \pm 4.73	71.98 \pm 2.58	77.39 \pm 1.70	68.04 \pm 2.59	42.85 \pm 5.04	66.46 \pm 2.17
GB	75.51 \pm 1.15	62.54 \pm 2.24	37.40 \pm 3.73	62.18 \pm 1.66	77.99 \pm 1.94	70.33 \pm 3.05	44.69 \pm 5.59	68.65 \pm 2.79
VE	79.87 \pm 1.77	72.97 \pm 2.51	50.29 \pm 4.74	70.99 \pm 2.34	77.35 \pm 2.85	70.83 \pm 2.70	44.31 \pm 6.47	69.41 \pm 2.28
SE	78.17 \pm 1.18	72.65 \pm 2.17	47.44 \pm 3.22	71.82 \pm 3.02	73.35 \pm 6.31	67.63 \pm 6.49	37.46 \pm 13.13	67.43 \pm 6.56

cases.

Table 3.6 presents the results of CBOW and skip-gram word embedding based features extracted from ASTD dataset. The highest results are obtained using CBOW with voting based ensemble classifier with slightly difference from skip-gram.

Table 3.7 presents the results of *tf-idf* and LSA features extracted from GS-dataset using the aforementioned classifiers. *tf-idf* archives higher results than LSA in most cases.

Table 3.8 presents the results of CBOW and skip-gram word embedding based features extracted from GS-dataset. The highest results are obtained using CBOW.

3.6.3 Oversampling Techniques Evaluation

Several experiments are conducted to evaluate the aforementioned over-sampling techniques on the considered datasets using SVM. It should be mentioned that, oversam-

Table 3.8: Word embedding based features performance using GS-dataset

Clf	CBOW				Skip-gram			
	Acc	F_1	MCC	GM	Acc	F_1	MCC	GM
SGD	85.06 \pm 1.57	81.89 \pm 2.18	65.23 \pm 4.07	81.24 \pm 3.54	85.06 \pm 1.36	81.90 \pm 2.07	65.18 \pm 3.45	81.30 \pm 3.49
SVC	85.55 \pm 1.72	82.02 \pm 2.00	65.28 \pm 4.13	80.28 \pm 1.92	85.87 \pm 1.68	82.34 \pm 2.29	65.98 \pm 4.20	80.55 \pm 2.45
GNB	74.97 \pm 1.78	70.49 \pm 2.21	41.20 \pm 4.36	70.38 \pm 2.38	74.26 \pm 1.95	69.31 \pm 2.62	38.92 \pm 5.09	69.05 \pm 2.79
NN	72.92 \pm 6.80	68.04 \pm 5.37	38.32 \pm 9.86	68.04 \pm 3.81	75.90 \pm 3.84	69.17 \pm 3.08	41.16 \pm 6.92	68.02 \pm 2.33
DT	72.38 \pm 2.85	68.04 \pm 2.84	36.30 \pm 5.61	68.22 \pm 2.71	70.75 \pm 2.06	66.01 \pm 2.37	32.06 \pm 4.74	66.09 \pm 2.40
RF	81.90 \pm 1.43	76.23 \pm 1.96	55.58 \pm 3.72	74.06 \pm 1.86	81.83 \pm 1.60	75.90 \pm 2.56	55.38 \pm 4.30	73.69 \pm 2.50
GB	84.17 \pm 1.26	80.28 \pm 1.60	61.68 \pm 3.07	78.65 \pm 1.66	83.92 \pm 1.91	79.82 \pm 2.80	60.97 \pm 5.03	78.18 \pm 3.08
VE	85.70 \pm 1.73	82.35 \pm 2.08	65.77 \pm 4.22	80.85 \pm 2.24	85.91 \pm 1.26	82.48 \pm 1.81	66.29 \pm 3.03	80.88 \pm 2.37
SE	85.55 \pm 1.71	83.15 \pm 2.23	66.74 \pm 4.04	83.38 \pm 2.86	85.38 \pm 1.20	82.26 \pm 2.03	65.67 \pm 3.24	81.53 \pm 3.27

Table 3.9: The effects of over-sampling techniques for ASTD

Technique	Acc	F_1	MCC	GM
Original	84.54 \pm 1.89	80.67 \pm 2.77	62.33 \pm 4.92	79.14 \pm 2.92
ROS	83.89 \pm 2.44	81.64 \pm 3.01	63.67 \pm 6.10	82.77 \pm 3.47
SMOTE	84.08 \pm 2.14	81.84 \pm 2.59	64.09 \pm 5.23	82.94 \pm 2.97
SMOTE-B1	83.34 \pm 2.03	81.27 \pm 2.30	63.17 \pm 4.62	82.83 \pm 2.50
SMOTE-B2	82.55 \pm 2.25	80.54 \pm 2.55	61.97 \pm 5.19	82.43 \pm 2.81
SMOTE-SVM	84.68 \pm 2.24	82.48 \pm 2.75	65.28 \pm 5.55	83.50 \pm 3.14
ADASYN	83.80 \pm 2.02	81.44 \pm 2.30	63.17 \pm 4.60	82.29 \pm 2.46

pling methods are only applied to the training sets since it is unreasonable to validate models using synthetic instances. Table 3.9 presents the results of ASTD without and with applying over-sampling techniques. The highest results are obtained after applying over-sampling technique SMOTE-SVM. Imbalanced-learn toolbox [161] are used to implement oversampling techniques. Table 3.10 presents the results of GS-dataset without and with applying over-sampling techniques. Applying over-sampling technique leads to improve F_1 , MCC and GM.

Further experiments are conducted to evaluate the effect of over-sampling techniques in case of higher imbalanced dataset. This is performed on ASTD with remov-

Table 3.10: The effects of over-sampling techniques for Gold-Standard dataset

Technique	Acc	F_1	MCC	GM
Original	85.87 \pm 1.68	82.34 \pm 2.29	65.98 \pm 4.20	80.55 \pm 2.45
ROS	84.98 \pm 2.01	82.88 \pm 2.08	66.01 \pm 4.05	83.66 \pm 1.87
SMOTE	85.09 \pm 2.32	83.08 \pm 2.46	66.45 \pm 4.77	84.01 \pm 2.22
SMOTE-B1	84.74 \pm 2.10	82.82 \pm 2.22	66.11 \pm 4.31	84.10 \pm 2.11
SMOTE-B2	84.42 \pm 2.25	82.55 \pm 2.34	65.69 \pm 4.48	84.03 \pm 2.11
SVM-SMOTE	85.13 \pm 2.30	82.93 \pm 2.50	66.03 \pm 4.92	83.48 \pm 2.38
ADASYN	85.09 \pm 1.91	83.03 \pm 2.00	66.35 \pm 3.85	83.89 \pm 1.92

Table 3.11: The effects of over-sampling techniques for higher imbalance ratio ASTD dataset

Technique	Acc	F_1	MCC	GM
Original	89.56 ± 1.72	61.65 ± 7.23	32.53 ± 15.42	58.78 ± 5.04
ROS	86.20 ± 2.28	74.01 ± 4.13	51.12 ± 8.67	81.57 ± 5.83
SMOTE	85.79 ± 2.36	73.53 ± 4.38	50.32 ± 9.19	81.33 ± 6.00
SMOTE-B1	86.44 ± 1.97	73.77 ± 3.66	49.92 ± 7.59	80.19 ± 4.90
SMOTE-B2	85.67 ± 2.57	72.62 ± 4.40	47.80 ± 9.03	79.10 ± 5.73
SMOTE-SVM	88.15 ± 2.15	74.86 ± 4.76	50.67 ± 9.82	78.56 ± 6.11
ADASYN	87.44 ± 2.07	74.13 ± 3.49	49.57 ± 6.87	78.37 ± 4.14

ing some examples from the minority class (positive). This leads to a dataset with 200 positive opinions with IR= 7.48.

3.6.4 Deep Learning Techniques Evaluation

Table 3.12 shows the results of the evaluated models. The highest results are presented in bold. In general, non-static models with the combined LSTMs give better results.

Different optimizers can be used to compile models on Keras² including: Adagrad, Adam, Rmsprop and SGD. The previous experiments were carried out using Adam optimizer [162]. Their impact on various models are investigated with their default parameters on ArTwitter dataset and non-static CBOW model. The average performance of Rmsprop is the best followed by Adam. Moreover, the highest results obtained for ArTwitter dataset is obtained using Rmsprop in case of the combined LSTMs with non-static word initialization model.

Finally, the highest attained performance is compared with that in the literature as shown in Table 3.14. It is clear that our proposed method of combining LSTMs compares favorably with other work.

Table 3.12: Performance comparison of various models on ASTD and ArTwitter datasets with static and non-static initializations for CBOW and skip-gram word embeddings

Word2vec	Dataset	Method	Static				Non-Static			
			Prec	Rec	Acc	F1	Prec	Rec	Acc	F1
CBOW	ASTD	CNN	74.86	74.40	74.40	74.43	74.12	74.10	74.10	74.11
		LSTM	75.04	74.70	74.70	74.74	80.12	80.12	80.12	80.07
		CNN-LSTM	71.18	68.07	68.07	67.58	76.92	73.49	73.49	72.00
		Stacked-LSTM	72.98	65.66	65.66	63.90	73.60	70.18	70.18	69.70
		Comined-LSTM-SUM	79.04	78.31	78.31	78.33	81.02	81.02	81.02	80.98
		Comined-LSTM-MUL	78.43	77.41	77.41	77.40	82.32	81.63	81.63	81.64
		Comined-LSTM-CONC	78.64	77.11	77.11	77.05	80.45	80.42	80.42	80.35
	ArTwitter	CNN	77.47	77.21	77.21	77.06	78.13	77.82	77.82	77.67
		LSTM	83.22	83.16	83.16	83.17	84.59	84.39	84.39	84.40
		CNN-LSTM	79.78	78.23	78.23	78.10	81.79	80.70	80.70	80.63
		Stacked-LSTM	82.54	82.34	82.34	82.35	82.12	81.93	81.93	81.85
		Comined-LSTM-SUM	82.58	82.55	82.55	82.55	84.80	84.80	84.80	84.80
		Comined-LSTM-MUL	83.01	82.96	82.96	82.96	85.42	85.42	85.42	85.42
		Comined-LSTM-CONC	83.22	82.96	82.96	82.96	86.46	86.45	86.45	86.45
Skip-grams	ASTD	CNN	73.96	61.45	61.45	57.5	73.96	66.57	66.57	64.90
		LSTM	76.85	76.51	76.51	76.54	77.88	77.41	77.41	77.44
		CNN-LSTM	76.35	75.90	75.90	75.56	75.34	71.99	71.99	71.58
		Stacked-LSTM	70.79	68.98	68.98	68.80	77.02	76.51	76.51	76.54
		Combined-LSTM-SUM	78.31	78.31	78.31	78.31	79.01	78.92	78.92	78.94
		Combined-LSTM-MUL	77.82	77.11	77.11	77.13	78.73	76.20	76.20	76.02
		Combined-LSTM-CONC	79.09	78.61	78.61	78.64	80.90	80.42	80.42	80.45
	ArTwitter	CNN	81.2	75.56	75.56	74.73	84.2	83.16	83.16	83.11
		LSTM	82.49	80.90	80.9	80.79	83.62	83.57	83.57	83.54
		CNN-LSTM	78.51	73.92	73.92	72.45	84.24	84.19	84.19	84.20
		Stacked-LSTM	82.21	81.72	81.72	81.72	82.95	82.96	82.96	82.95
		Combined-LSTM-SUM	83.04	82.55	82.55	82.54	85.64	85.63	85.63	85.61
		Combined-LSTM-MUL	82.28	81.72	81.72	81.71	85.83	85.83	85.83	85.82
		Combined-LSTM-CONC	81.45	81.31	81.31	81.32	87.36	87.27	87.27	87.28

Table 3.13: Compilation optimizers with ArTwitter and non-static CBOW model

Method	Adagrad		Adam		Rmsprop		SGD	
	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1
CNN	52.36	35.99	77.82	77.67	78.85	78.84	79.67	79.64
LSTM	85.83	85.84	84.39	84.40	84.19	84.19	68.79	68.77
CNN-LSTM	80.29	80.24	80.70	80.63	82.75	82.72	82.34	82.30
Stacked-LSTM	84.19	84.19	81.93	81.85	84.19	84.19	57.49	53.89
Combined-LSTM-SUM	84.80	84.81	84.80	84.80	83.37	83.38	66.74	66.30
Combined-LSTM-MUL	85.01	85.02	85.42	85.42	86.65	86.65	64.07	64.08
Combined-LSTM-CONC	86.04	86.04	86.45	86.45	87.06	87.07	65.71	65.72
Average	79.79	77.45	83.07	83.03	83.87	83.86	69.26	68.67

Table 3.14: Comparisons with other related approaches

Dataset	Approach	Technique	Accuracy
ASTD	Dahou et. al [78]	CNN non-static	75.90
	Our work	Combined-LSTM-Mul, non-static, CBOW, Adam optimizer	81.63
ArTwitter	Dahou et. al [78]	CNN non-static	85.01
	Abdulla et. al [50]	Root-stemmer + SVM	85.00
	Our work	Combined-LSTM-CONC, non-static, Skip-gram, Adam optimizer	87.27

3.7 Summary

In summary, for comparing traditional and word embedding based features, the experiments revealed:

- Word embedding based features achieve the highest results comparing to the traditional features in all cases.
- Voting-based ensemble classifier achieve the best results in most cases which is followed by SVM.

Regarding applying and evaluating oversampling technique with word embeddings, the experiments revealed:

- SMOTE-SVM achieves the best performance, for ASTD dataset, in terms of all considered evaluation measurements. Other oversampling techniques leads to improve the results except SMOTE-B2.
- For Gold-standard dataset SMOTE achieves the best performance in terms of F_1 and MCC. The highest recall and GM are obtained using SMOTE-B1. Accuracy is dropped when applying oversampling techniques.

- Oversampling techniques improve the results significantly in terms of F_1 , MCC and GM with higher imbalance ratio dataset.

Regarding deep learning approach, the experiments revealed:

- Deep learning models based on word2vec vectors updated during learning achieves the highest results in nearly all cases.
- LSTM based models perform better than CNN. This can be attributed as that LSTMs perform well with sequences of data in our case the sentences while CNNs don't have the capability to understand the context of sentences. With LSTMs each word is processed based on the understanding of the previous words.
- The proposed combined LSTM architectures perform better than other models.

CHAPTER 4

TEXTUAL-EMOJIS SENTIMENT ANALYSIS

The aim of this chapter is at alleviating issues of text-based sentiment analysis such as domain-, topic-, temporal-independent through incorporating different sources such as visual modality (emojis) and evaluating their effectiveness to detect polarities in microblogs. First a dataset for Arabic microblogs, in which each instance contains at least an emoji, is prepared. Different feature representations for emojis are proposed and evaluated, including: emoji embeddings, emoji frequencies and lexicon-based emojis features. The impact of emojis on text for sentiment classification of dialectical Arabic is analyzed using several proposed fusion schemes. An approach is also presented to address emojis imbalance problem based on bagging algorithm and oversampling methods.

4.1 Background

Emoticons and emojis are first defined and distinguished in the following subsection.

Then, we review emojis in social media and provide a taxonomy of related studies.

Lastly, we survey approaches that applied emojis for the sentiment analysis task.

4.1.1 Emoticons vs. Emojis

Emoticon is an abbreviation of “emotion icon” and is considered as ASCII character sequence (not in image) such as :), :(, ;-), etc. Each represents a facial expression of the authors’ feelings in a written matter as a succession of characters with non-verbal elements. Although, emoticons are defined in the computer-mediated communication literature in different words, they have the same meaning and semantic. Emoticons are referred as “relational icons” in [163]. They are defined as “a sequence of ordinary characters you can find on your computer keyboard. Smileys are used in e-mail and other forms of communication using computers” in [164]. In [165], emoticons serve different functions in digital conversations and categorized them into five different classes, namely: “emotion icons”, “social markers of familiarity”, “pragmatic markers”, “structural markers” and “creative resources”.

Emoticons are divided into Western and Eastern emoticons. Western emoticons are the most popular and most frequently used such as :), =), xD, etc. They are usually horizontal. This type is known as *emoticon* and has a limited representativeness [166]. On the other hand, Eastern emoticons are known as *kaomoji* and are usually vertical such as ((+_+)) and (+O+) for confusion, (-_-)zzz for sleeping, (= _ =) for

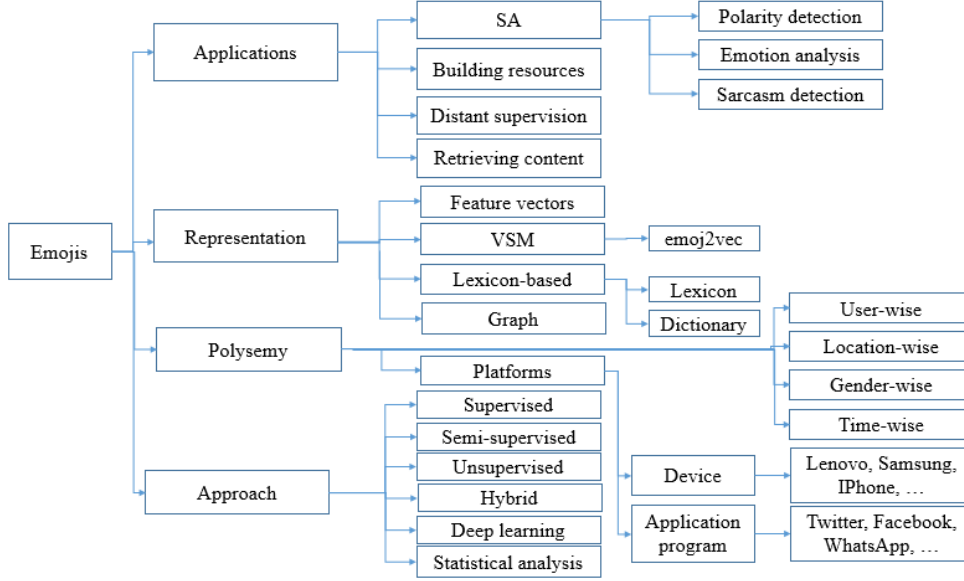


Figure 4.1: Emojis taxonomy

tiredness, (-_-)!! and (-.-) for shame and _U~~ represents a cup of tea. Eastern emoticons are able to represent more complex faces and body positions than western emoticons [166]. Due to the nature of Arabic writing from right-to-left, smileys and sad emoticons might be mistakenly interchanged [167, 168].

By contrast, emojis are recognized as successor to emoticons. They do not just represent facial expressions but also several roles and relations such as fun elements for occasions, objects, travels, food and drink, animals, countries and activities [169]. In contrast to emoticons, emojis are represented by images.

4.1.2 Emojis in Social Media

A taxonomy shown in Figure 4.1 in which prior works on emojis are classified according to their applications, representations, issues, and approaches.

A) Applications

Emojis are applied to build different resources including emoji-embedding models [170, 171], lexicons [172, 173], datasets, etc. Supervised machine learning sentiment analysis approaches require highly annotated dataset which is tedious, labor-intensive and time consuming. Adopting emoticons/ emojis for annotating training sentiment analysis datasets leads to alleviating this issue [174]. However, it is not recommended to use whole training dataset annotated automatically for building classifier because the noise in data. Therefore, Liu et al. [175] proposed a model called emoticon smoothed language model (ESLAM) for utilizing and smoothly integrating both manually and noisy labeled data for building training subset. The ESLAM first utilizes manually labeled data to train language models and then the noisy labeled data is utilized for smoothing. A distant supervision approach [176] is an application to utilize emoticons/ emojis, hashtags, etc, as noisy labels to automatically annotate training datasets for different sentiment analysis tasks [168, 177, 178]. Another task is to predict the most likely emoji given the text of a tweet [179–181].

B) Representations

First, it is worth mentioning that there are significant differences between “Representation” and “Appearance” concepts, in this study. “Representation” means how emojis are represented to be addressed as a concern/research problem or to be employed for different tasks. “Appearance” means how emojis look/appear for users and it might change from platform to another. Emojis can be represented in different forms depending on the application or task. They are mapped into a representation

of characteristics (features) as discriminators. With such type of representation emojis are mapped into a set of values. These values might be binary numbers (0 or 1) such as the existence of emoji in an instance, integer numbers when taking the count of emojis in instances, or real number. Similar to term, word or phrase polarity lexicons, emojis have lexicons in which each emoji has a polarity or sentiment score. Emoji Sentiment Ranking (ESR) is a systematic lexicon of emojis built for sentiment analysis by Novak [172]. It is composed of 969 emojis; 751 of them occur greater than four times. Each emoji is assigned a sentiment score computed from 1.6 million tweets in 13 European languages by the sentiment polarity (negative, neutral, or positive). Another emoji sentiment lexicon with 840 emojis using an unsupervised sentiment analysis system was constructed by [182]. It was built based on the definitions given by emoji creators in Emojipedia while analyzing the sentiment of informal texts in English and Spanish. Moreover, lexicon variants were created by considering the sentiment distribution of the informal texts accompanying emojis.

Emoji embedding models are other representations of emojis, in which each emoji is represented as a vector of real numbers generated using well-known embedding tools such as word2vec [66, 67]. Emojis are represented as embeddings which can be readily used in downstream social natural language processing applications.

C) Polysemy

Several factors led to the ambiguity in interpreting emojis, e.g. users [183–185], gender [186–188], locations [189, 190], cultures [191], platforms [183, 192, 193], etc. It was reported in [183, 192] that there are significant variations between people’s in-

Apple	Facebook	Google	Twitter	Windows
				

Figure 4.2: Example of “cow face ” emoji appearance on different platforms

terpretation of emoji ratings within and across platforms. An emoji might appear differently in various platforms and devices such as Twitter, Facebook, Apple, Emojione, Samsung, etc. An emoji can also appear differently in different versions of the same system such as Android. Figure 4.2 shows an example for the emoji of “cow face” appearance in different platforms based on version 11.0 of Full Emoji List [194]. Another major factor that causes misunderstanding is the similarity of different emojis such as octopus 🐙 and squid 🐙. Moreover, users may be unfamiliar with some emojis such as the use of “pile of poo” 💩 which has negative polarity with sentiment score of -0.116 in ESR and they misuse it to represent “ice cream” which is definitely positive with sentiment score of 0.212 in ESR.

Emoji sentiment perception from writers to readers viewpoints is another factor. Berengueres and Castro [195] reported that there is an 82% agreement in emoji sentiment perception from writers to readers viewpoints. The disagreement concentrates in negative emojis, where authors report to feel 26% worse than perceived by readers. Emoji usage was not found to be correlated with author’s moodiness. Emoji sentiments are interpreted in a different way according to the platform. It was concluded in [183] that there is disagreement in sentiment and semantics of 22 emojis on 5 different platform renderings especially across platforms. On the other hand, Cui et al. [196] studied the use of tweets whose sentiments conflict to some extent to emojis

in training phase. The main findings were that the optimal training dataset for determining tweets' sentiment is reasonable and followed the distribution of sentiment in real tweet streams. Users tend to use emojis with positive polarity or happy emotion more than other polarities or emotions [195]. Additionally, young people tend to use emojis more frequently [185, 197].

Emojis are analysed and studied in different social media platforms including: Twitter such as [170, 198], Facebook [199, 200], WhatsApp [185], Instagram [181], electronic mail [201], etc. Furthermore, the use of emojis on different social media including WhatsApp, Facebook and Twitter are analysed in [169]. The main findings were that most popular Emojis in one social media is not as popular in the others. Emojis sentiment polarity in Twitter is high and overall number of Emojis is less than Facebook. The sentiment value of Emojis is more meaningful when there are multiple Emoji in one notification.

Significant attempts and efforts have been performed to clarify the meanings and to reduce arisen misunderstandings. For example, Wijeratne et al. [202, 203] presented the first and largest machine readable sense inventory for emoji (EmojiNet). EmojiNet links Unicode emoji representations to their English meanings extracted from the Web. It is composed of a dataset of 12,904 sense labels over 2,389 emoji. Each emoji sense is associated with context words trained using skip-gram word2vec technique.

D) Approaches

Emojis-based studies can be classified into machine learning [196, 204] and statistical analysis studies. Barbieri et al. [170] built several skip-gram embedding models using

a dataset of 10 millions tweets by mapping in the same vectorial space both words and emojis. The tweets were posted by USA users. The models were then evaluated with semantic similarity experiments and compared with human assessment. Another emoji embedding model was trained in [198], using skip-gram word2vec technique in pre-trained embeddings for all Unicode emojis which are learned from their description in the Unicode emoji standard. This model outperforms a skip-gram model trained on a large collection of tweets. Barbieri et al. [179] trained several supervised classifiers based on deep learning, Long Short-Term Memory networks, for predicting appropriate emojis from corresponding tweets. The main conclusion was that computational models can identify the underlying semantics of emojis better than humans do. Statistical analysis studies were conducted to analyze the behavior of emoji [188, 191, 205, 206]. For example, [188] conducted a statistical analysis to explore the emoji usage through males and females in terms of the frequency, preferences, input patterns, public/private Computer-Mediated Communication-scenario patterns, temporal patterns, and sentiment patterns. They found that males and females varied in emoji usage significantly which confirms the findings of [187]. Another statistical analysis is conducted to investigate the functions of emojis from the perspective of the original senders by [206]. The main finding was that the social and linguistic function of emojis are complex and varied. It was reported in [205] that Twitter users tend to reduce their usage of emoticons and shift dramatically to use emojis.

4.1.3 Emojis in Sentiment Analysis

Emoticons or limited numbers of emojis were considered without taking into account the extent of the sentiment representation they convey [166]. Emoticons only have been considered as elementary/extra features for sentiment analysis tasks such as the number of negative or positive emoticons [207, 208] or the presence of positive and/or negative emoticons [48, 168, 208]. In addition, emoticons were converted to their textual meanings, as a preprocessing step, intuitively or using a general emoticons lexicon [207, 209–211]. The presence and the count of emoticons/emojis, the number of positive and negative emoticons, as features, were considered and evaluated to detect emotion in tweets in [212].

Some attempts have considered the construction of emoji-related resources for NLP tasks such as datasets, lexicons, dictionaries and even tools. Emojis’ lexicons have been constructed for sentiment analysis tasks. Hogenboom et al. [173] presented a lexicon-based polarity classification method to evaluate how emoticons convey sentiment. This method was evaluated on 2,080 Dutch tweets and forum messages, which all contain emoticons. They reported that the sentiment of emoticons tends to dominate the sentiment conveyed by textual cues and forms a good proxy for detecting the polarity of text.

4.2 Emojis-based Sentiment Classification

This section describes the main operations performed to achieve our objective of this chapter. The layout of the proposed sentiment classification approach for single modal-

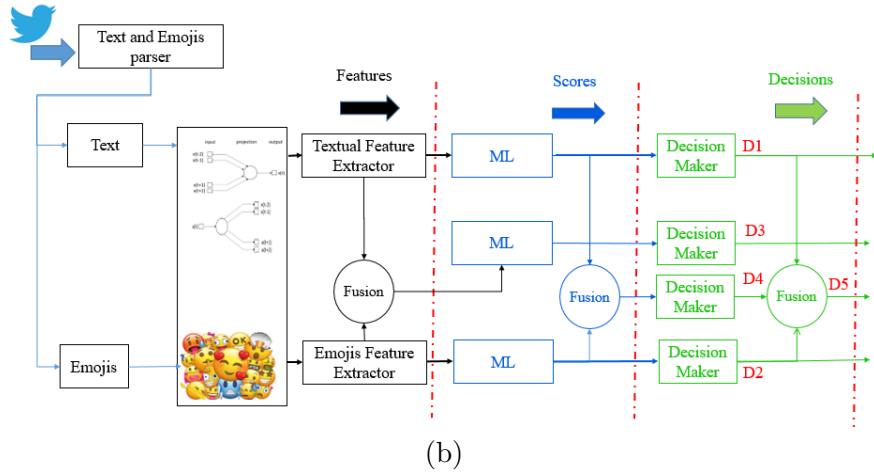
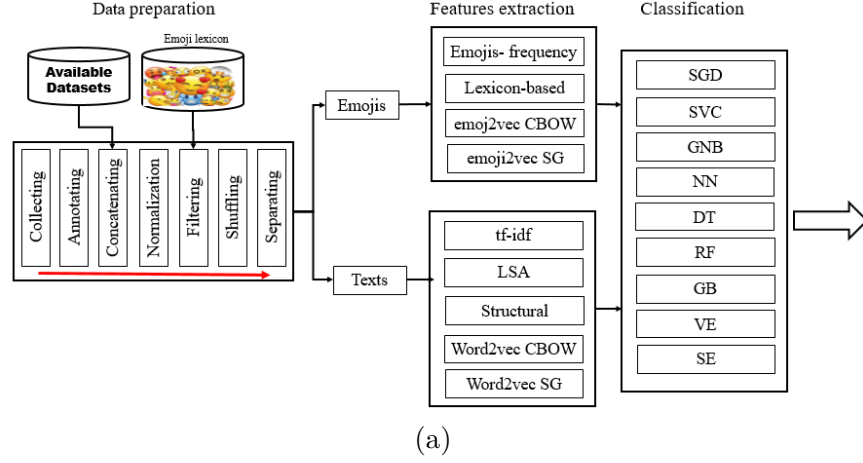


Figure 4.3: A general overview of the proposed approach (a) the basic model of single modalities of text and emojis (b) the fusion model of both modalities at different levels

ities and fused modalities is illustrated in Figure 4.3.

4.2.1 Dataset Preparation

Machine learning based approach requires annotated datasets which are often difficult or even impossible to obtain. This is mainly because labeling data takes considerable human effort. Therefore, the used dataset is prepared from five publicly available sentiment related datasets [49–51, 167, 213] as well as additional microblogs collected as described in Table 4.1. Instances having at least one emoji of an existing lexicon of

Table 4.1: Description of the evaluated dataset showing the various sources and number of instances that contain emojis

Source	Instances having emojis		
	Negative	Positive	Total
Existing datasets: [49–51, 167, 213]	462	786	1248
Additional instances:	413	430	843
Total	875	1216	2091

emojis [172] are used. All instances were manually annotated as positive or negative. First of all, emoticons expressed in ASCII encoding were normalized and transformed to their corresponding graphical symbols. A total of 2091 instances distributed as 1216 positive and 875 negative is finally kept. The reason of having more positive instances than negative instances is our observation that users tend to frequently use emojis when they are happy. This confirms the findings reported by [195]. Finally, the instances are shuffled randomly. Due to the issues related to emojis’ misunderstanding and ambiguity issues, the dataset used is prepared carefully to alleviate these issues. The sentiment of each instance was annotated based on the text without considering emojis.

Preprocessing step is conducted at three levels relying on the type of features. For emoji based features, as mentioned ASCII emoji codes are transformed into their corresponding graphical symbols. For tweet based features, no preprocessing operation is conducted since these features are sensitive to content. For textual features, different operations are conducted on the text including: Removing noisy symbols, non-Arabic characters, diacritical marks, punctuation marks, links, and repeated characters.

4.2.2 Emojis-based Features

Several emojis based features are considered and investigated to be used as main discriminators for sentiment classification task, including emojis frequency, lexicon based features and two forms of emoji2vec embedding based features. To our knowledge, this is the first time to investigate those features to determine the sentiment in Arabic microblogs.

A) Emojis frequency

The count of occurrences of each emoji in each instance in the prepared dataset is calculated. Feature vectors are prepared for the 2091 instances of dimension 429. Since the aim is to investigate emojis on Arabic sentiment analysis, there is a need to evaluate how each emoji is important or significant to predict sentiment polarity. Two popular feature selection techniques namely: ReliefF and Correlation-Attribute Evaluator (CAE) are applied. These two algorithms for feature ranking are fast and require linear time in the number of features and instances. The ReliefF method computes a weight for each emoji by sampling an instance repeatedly and taking into account the value of the given attribute for the closest sample of the same and different class [214]. On the other hand, the CAE method computes the Pearson's correlation of the emoji and the polarity label. Figure 4.4 (a) and (b) show the top-ranked 30 emojis by ReliefF and CAE, respectively for emojis frequency features. It is found that 19 out of the 30 top ranked emojis are shared in both figures.

B) Lexicon-based features

ESR lexicon is employed in order to extract features. It is three-class polarity (negative, neutral and positive) lexicon. It was constructed using a dataset of around 1.6 million tweets expressed in 13 European languages. The tweets were annotated as negative (-1), neutral (0) or positive (+1) by 81 annotators. Therefore, an emoji takes the sentiment from the tweet where it appears. ESR is composed of 969 emojis, 751 of them occur more than five times in the dataset. A discrete probability distribution for each emoji is computed p_- , p_0 , p_+ for emoji's negativity, neutrality and positivity, respectively. The probability p_c for an emoji is calculated as:

$$p_c = N_c/N \quad (4.1)$$

where c is the label, $c \in \{-1, 0, 1\}$, N_c the number of the emoji's occurrences in tweets with label c , and N is the number of the emoji's occurrences in all tweets. In case of computing the probability sentiment distributions, it was considered that an emoji might occur in a tweet many times. Then sentiment score (ss) of each emoji was calculated by subtracting the negativity probability from the positivity,

$$ss = p_+ - p_- \quad (4.2)$$

Figure 4.5 shows the top-10 emojis appear in ESR. The feature vectors are extracted based on the scores defined on the aforementioned lexicon. The feature i in tweet j ,

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)
😂		0x1f602	14622	0.805	0.247	0.285	0.468	0.221	
♥		0x2764	8050	0.747	0.044	0.166	0.790	0.746	
♥		0x2665	7144	0.754	0.035	0.272	0.693	0.657	
😍		0x1f60d	6359	0.765	0.052	0.219	0.729	0.678	
😭		0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093	
😘		0x1f618	3648	0.854	0.053	0.193	0.754	0.701	
😊		0x1f60a	3186	0.813	0.060	0.237	0.704	0.644	
👉		0x1f44c	2925	0.805	0.094	0.249	0.657	0.563	
💕		0x1f495	2400	0.766	0.042	0.285	0.674	0.632	
👋		0x1f44f	2336	0.787	0.104	0.271	0.624	0.520	

Figure 4.5: Top 10 emojis in ESR

f_{ij} is computed as:

$$f_{ij} = ef_{ij} * ss_i \quad (4.3)$$

where ef_{ij} is the emoji frequency of emoji i in tweet j while ss_i is the sentiment score of the emoji i in the lexicon.

C) Emojis Embedding

As mentioned in the previous chapters, embedding techniques are recognized as an efficient method for learning high-quality vector representations of words, terms or phrases from large amounts of unstructured text data. They refer to the process of mapping words, terms or phrases from the vocabulary to real-valued vectors such that elements with similar meanings to have similar representations. In this chapter, word2vec technique is used to map emojis to real-valued vectors and called *emoji2vec*. Emojis embeddings are generated based on employing CBOW and skip-grams neu-

Table 4.2: Training parameters emojis embeddings

Model	Dimensionality	Window size	Sample	Negative	Min count.	Iterations
CBOW/SG	300	5	$1 \times e^{-3}$	10	10	10

ral embedding techniques to be used as main features. Emojis are mapped into d -dimensional embeddings. The dataset prepared in [215] of one million sentences collected from Twitter each contain emojis is utilized to generate the emojis embedding models. The parameters used to generate emoji2vec models are depicted in Table 4.2.

Several semantics and syntactical relations can be obtained using emoji2vec. For example, Figure 4.6 shows four different information types can be obtained using emoji2vec. For each query, the 10 highest probability answers, or less based on the availability, are retrieved. They are ordered based on their probabilities from left to right and from up to down. The first query is to retrieve the relation (king+man-woman) with the 10 highest probability using CBOW and skip-gram models. Both techniques agree in the first answer which is 👑 crown. This is similar to the well-known example of vector of (King) - the vector of (Man) + the vector of (Woman) is close to the vector of (Queen). CBOW and skip-grams differ in the order of some other retrieved answers while others are common. The second query is to retrieve the most related emojis for a concept or word. The figure depicts three examples for this query. Another query is to retrieve the most likelihood emojis related semantically to a certain emoji and two examples are shown in the same figure. The last query is to estimate the similarity of two emojis.

Now, assume a tweet T of n emojis after filtering words, $T = \{e_1, e_2, \dots, e_n\}$,




























































Query	CBOW	Skip-grams
King + woman – man	    	    
Marriage	  	  
Hate	    	   
Love	    	    
	    	    
	    	    
 & 	0.563	0.444

Figure 4.6: Different queries for emoji2vec using CBOW and skip-grams

where e_i is the i^{th} emoji in T . Let $x_i \in \mathbb{R}^d$ be the d -dimensional emoji vector corresponding to the i^{th} emoji in T . To compute the feature vector for T , the feature vectors of emojis are arranged in a matrix column-wise then the row-wise average is computed as illustrated in Figure 4.7 to obtain the feature vector:

$$f_i = \frac{1}{n} \sum_{k=1}^n x_{k,i}, i = 1, 2, \dots, d \quad (4.4)$$

4.2.3 Fusion with Textual Features

Since it is the first time to evaluate such features, it is essential to draw and build the baseline to compare the capability of the proposed features. Towards this end, we compared the performance of them with several textual features using the same experimental settings. For textual features, $tf-idf$, LSA, structural features and word embeddings are adopted. $tf-idf$, LSA and word embedding features are described in



Figure 4.7: An example of sentence representation using emoji2vec embeddings

Chapter 3. The followings are considered as structural features:

- Count of links: its value is equal to the number of links in the tweet otherwise zero value is assigned.
- Count of mentioned accounts: if a tweet mentions Twitter's accounts, the number is assigned otherwise it takes zero value.
- Count of hashtags: if a tweet contains hashtags, the count of hashtags is assigned otherwise it takes zero value.
- Count of emojis
- Is tweet with elongation words?: elongation words mean some characters are repeated such as Nooooo! Hiiiiiiiiiii!.
- Is tweet with diacritical marks?
- Length of tweet in words.

- Length of tweet in characters.

For word embedding based features, recent pretrained models generated using the CBOW and skip-grams are adopted [216]. The models were learned using a corpus of more than 77,600,000 Arabic tweets posted between 2008 and 2016. Tweets were written in MSA and different Arabic dialects. A dimensionality of 100 and a window size of three were used for generating both models. A feature vector with size of 100 attributes is created by calculating the average of the embedding vectors of that sample as described in Section 3.2.3. The most discriminating features extraction methods are then selected for the remaining experiments. The effect of combining emojis with textual features in different fusion levels are explored and investigated, including feature, score and decision levels.

A) Feature-level fusion

Feature-level fusion is carried out through simply concatenating the extracted features including textual and emojis. Mathematically, let $F = \{f_1, f_2, \dots, f_n\}$, $F \in R^n$ represents the textual feature vector with length of n extracted from a tweet and $E = \{e_1, e_2, \dots, e_m\}$, $E \in R^m$ represents emojis feature vector with size of m . Combining F and E is the goal which results in a new feature vector, $C = \{f_1, f_2, \dots, f_n, \dots, e_1, e_2, \dots, e_m\}$, $C \in R^k$, with size of k such that $k = n + m$. Two inherent issues arise when fusing features and need to be addressed, namely scaling and the curse of dimensionality. The former issue is due to the feature values extracted using different methods are scaled differently. Moreover, some features might be redundant or noisy. These two issues are handled through normalization and features selection.

Features are normalized using MINMAX scheme to produce $F_{norm} = \{f'_1, f'_2, \dots, f'_n\}$, $E_{norm} = \{e'_1, e'_2, \dots, e'_m\}$, respectively:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4.5)$$

For feature reduction, PCA is applied with the criteria of select the number of components such that the amount of variance that needs to be explained is greater than 0.99

B) Score-level fusion

Assume that there is k matchers, $\{M_1, \dots, M_k\}$, s_i is the score of matching the features of an instance $\{x_1, \dots, x_n\}$ using matcher M_i , then the overall score (S) can be computed using several popular schemes, including:

$$\text{Max rule, } S = \max\{s_1, s_2, \dots, s_k\} \quad (4.6)$$

$$\text{Sum rule, } S = \sum_{i=1}^k s_i \quad (4.7)$$

$$\text{Product rule, } S = \prod_{i=1}^k s_i \quad (4.8)$$

The scores are normalized using MINMAX scheme (Eq. 4.5).

Three possibilities are conducted to combine different modes in the score level, including: (1) CBOW and Emojis, (2) skip-grams and Emojis, (3) CBOW, skip-grams and Emojis

Decision-level fusion

In case of having k decision makers (models) $\{DM_1, \dots, DM_k\}$, such that d_i is the decision made by DM_i for an instance $\{x_1, \dots, x_n\}$, the general decision can be made from the local decisions using several methods. We adopt the voting method, in which the final decision is equal to the most frequent decision made using the local decisions:

$$\hat{y} = mode\{d_1, d_2, \dots, d_n\} \quad (4.9)$$

4.3 Experiments

Different classifiers are trained to evaluate the proposed features in order to classify the sentiment. The used classifiers are described in Table 3.3. The experiments are evaluated using 10-fold cross validation method. Our experiments are conducted in Python utilizing scikit-learn package [160] and imbalanced-learn toolbox [161].

4.3.1 Results and Discussion

Table 4.3 shows the results obtained using the considered five textual features: *tfidf*, LSA, structural features, CBOW and skip-grams. For *tf-idf*, the best performances are obtained using LR classifier. For LSA the highest results are obtained also using LR classifier. The lowest results are obtained using structural features. In general, for textual features skip-gram achieves the highest results which is followed by CBOW.

Emojis-based results for all considered machine-learning approaches are reported in Table 4.4. Four feature extraction methods are evaluated: emojis frequencies,

Table 4.3: Performance comparison of ten machine-learning approaches using textual features extracted by five different methods; the highest results are marked in bold.

clf	Recall	Prc	F_1	GM	Acc
Feature Extraction Method: $tf-idf$					
SGD	70.68 \pm 2.77	73.48 \pm 1.92	70.39 \pm 2.93	71.29 \pm 2.61	70.68 \pm 2.77
SVC	72.55 \pm 2.11	73.59 \pm 1.71	72.63 \pm 2.13	72.74 \pm 1.92	72.55 \pm 2.11
GNB	66.57 \pm 3.11	70.51 \pm 3.21	66.49 \pm 3.19	68.44 \pm 3.06	66.57 \pm 3.11
NN	64.04 \pm 2.64	64.76 \pm 2.23	64.19 \pm 2.64	63.84 \pm 2.36	64.04 \pm 2.64
LR	74.41 \pm 2.31	75.61 \pm 1.84	74.49 \pm 2.29	74.75 \pm 2.01	74.41 \pm 2.31
DT	66.43 \pm 1.79	67.86 \pm 1.79	66.59 \pm 1.81	66.87 \pm 1.84	66.43 \pm 1.79
RF	72.17 \pm 2.78	72.92 \pm 2.51	72.29 \pm 2.75	72.21 \pm 2.61	72.17 \pm 2.78
GB	67.72 \pm 1.87	70.70 \pm 3.10	63.87 \pm 2.20	62.56 \pm 1.99	67.72 \pm 1.87
VE	72.88 \pm 2.20	75.13 \pm 1.51	72.75 \pm 2.31	73.40 \pm 2.12	72.88 \pm 2.20
SE	72.41 \pm 2.40	72.68 \pm 2.27	72.04 \pm 2.37	71.02 \pm 2.29	72.41 \pm 2.40
Feature Extraction Method: LSA					
SGD	70.30 \pm 3.83	72.56 \pm 3.05	69.96 \pm 4.03	70.47 \pm 3.53	70.30 \pm 3.83
SVC	70.83 \pm 2.48	71.98 \pm 2.85	70.94 \pm 2.53	71.10 \pm 2.93	70.83 \pm 2.48
GNB	64.56 \pm 1.96	67.24 \pm 2.13	64.66 \pm 1.97	65.81 \pm 2.06	64.56 \pm 1.96
NN	68.87 \pm 2.62	68.72 \pm 2.67	68.59 \pm 2.80	67.46 \pm 2.97	68.87 \pm 2.62
LR	71.78 \pm 2.51	72.22 \pm 2.70	71.86 \pm 2.55	71.50 \pm 2.83	71.78 \pm 2.51
DT	63.89 \pm 2.59	64.19 \pm 2.19	63.92 \pm 2.47	63.16 \pm 2.21	63.89 \pm 2.59
RF	70.11 \pm 2.61	70.07 \pm 2.44	69.98 \pm 2.52	68.99 \pm 2.38	70.11 \pm 2.61
GB	70.35 \pm 3.20	70.22 \pm 3.35	69.83 \pm 3.20	68.41 \pm 3.14	70.35 \pm 3.20
VE	71.30 \pm 3.61	72.63 \pm 2.98	71.22 \pm 3.59	71.29 \pm 3.13	71.30 \pm 3.61
SE	65.28 \pm 8.05	66.83 \pm 7.22	63.86 \pm 9.50	64.66 \pm 7.97	65.28 \pm 8.05
Feature Extraction Method: Structural Features					
SGD	53.30 \pm 7.23	52.31 \pm 13.40	43.13 \pm 11.55	52.12 \pm 3.33	53.30 \pm 7.23
SVC	55.90 \pm 4.59	42.80 \pm 14.15	41.95 \pm 3.31	50.30 \pm 1.61	55.90 \pm 4.59
GNB	58.21 \pm 3.30	61.49 \pm 3.55	58.12 \pm 3.39	59.77 \pm 3.40	58.21 \pm 3.30
NN	55.96 \pm 3.20	56.42 \pm 2.86	56.05 \pm 3.00	55.19 \pm 2.87	55.96 \pm 3.20
LR	59.16 \pm 2.82	61.28 \pm 2.54	59.32 \pm 2.85	59.99 \pm 2.63	59.16 \pm 2.82
DT	58.74 \pm 2.46	58.92 \pm 2.42	58.74 \pm 2.39	57.75 \pm 2.43	58.74 \pm 2.46
RF	58.45 \pm 2.85	57.96 \pm 2.92	58.07 \pm 2.90	56.63 \pm 2.96	58.45 \pm 2.85
GB	60.89 \pm 2.02	60.50 \pm 2.01	60.50 \pm 2.02	59.15 \pm 2.01	60.89 \pm 2.02
VE	53.78 \pm 6.77	57.67 \pm 10.22	44.47 \pm 10.20	52.54 \pm 3.10	53.78 \pm 6.77
SE	55.06 \pm 6.24	36.94 \pm 6.99	40.07 \pm 5.52	49.35 \pm 0.48	55.06 \pm 6.24
Feature Extraction Method: word2vec-CBOW					
SGD	81.20 \pm 3.17	82.40 \pm 2.58	81.18 \pm 3.25	81.43 \pm 3.17	81.20 \pm 3.17
SVC	81.49 \pm 2.43	82.10 \pm 2.51	81.58 \pm 2.43	81.76 \pm 2.59	81.49 \pm 2.43
GNB	74.32 \pm 2.77	76.27 \pm 3.08	74.47 \pm 2.74	75.33 \pm 3.00	74.32 \pm 2.77
NN	73.84 \pm 2.71	74.10 \pm 2.43	73.85 \pm 2.57	73.27 \pm 2.22	73.84 \pm 2.71
LR	81.44 \pm 2.50	82.22 \pm 2.57	81.54 \pm 2.50	81.84 \pm 2.63	81.44 \pm 2.50
DT	71.16 \pm 3.69	71.21 \pm 3.70	71.15 \pm 3.67	70.36 \pm 3.74	71.16 \pm 3.69
RF	80.10 \pm 2.69	80.12 \pm 2.73	79.86 \pm 2.75	78.70 \pm 2.87	80.10 \pm 2.69
GB	80.96 \pm 3.32	80.99 \pm 3.34	80.89 \pm 3.38	80.21 \pm 3.62	80.96 \pm 3.32
VE	81.63 \pm 3.08	82.39 \pm 2.99	81.71 \pm 3.09	81.98 \pm 3.18	81.63 \pm 3.08
SE	80.73 \pm 3.46	81.22 \pm 3.48	80.58 \pm 3.66	80.08 \pm 4.21	80.73 \pm 3.46
Feature Extractor Method: word2vec Skip-Gram					
SGD	81.54 \pm 3.10	82.76 \pm 2.37	81.54 \pm 3.12	81.83 \pm 2.79	81.54 \pm 3.10
SVC	82.64 \pm 3.03	83.31 \pm 3.12	82.73 \pm 3.02	83.02 \pm 3.22	82.64 \pm 3.03
GNB	74.99 \pm 2.79	76.78 \pm 3.05	75.14 \pm 2.77	75.93 \pm 2.98	74.99 \pm 2.79
NN	72.21 \pm 3.70	72.34 \pm 3.72	72.13 \pm 3.63	71.31 \pm 3.58	72.21 \pm 3.70
LR	82.54 \pm 2.64	83.41 \pm 2.90	82.64 \pm 2.64	83.04 \pm 2.91	82.54 \pm 2.64
DT	69.53 \pm 3.76	69.77 \pm 3.77	69.57 \pm 3.77	68.96 \pm 3.93	69.53 \pm 3.76
RF	80.97 \pm 2.25	81.10 \pm 2.16	80.73 \pm 2.35	79.65 \pm 2.56	80.97 \pm 2.25
GB	82.93 \pm 2.22	82.93 \pm 2.26	82.89 \pm 2.24	82.29 \pm 2.46	82.93 \pm 2.22
VE	82.64 \pm 3.47	83.44 \pm 3.48	82.71 \pm 3.45	82.98 \pm 3.57	82.64 \pm 3.47
SE	81.96 \pm 3.40	82.59 \pm 3.20	81.91 \pm 3.45	81.70 \pm 3.67	81.96 \pm 3.40

lexicon-based features, emojis CBOW and Skip Gram models. Compared to the well-known textual features namely *tf-idf*, LSA and structural features, the basic form of emojis based features with similar machine learning approaches, i.e. emojis frequencies models performs significantly better. The highest performance of emojis based features is achieved when using LR classifier with Skip Gram features, reaching an accuracy of $79.53\% \pm 2.03$. Although it is lower by 3.11% than the best approach for textual features, it has lower computational complexity than the extraction method of textual features.

The results obtained from textual based features and emojis features (presented in Tables 4.3 and 4.4) demonstrate that Word2Vec Skip Gram achieves the highest results followed by Word2Vec CBOW and then emojis. These individual feature extraction approaches are considered here as baseline uni-modal predictive models of tweets sentiment. We ran several experiments to evaluate different early and late fusion methods to improve the results. We considered fusing emojis features in their basic form (emoji frequencies), which only requires counting, with textual features using Word2Vec CBOW and Word2Vec Skip Gram. Table 4.5 illustrates the attained results for feature-level, score-level and decision-level fusions of two and three feature representations (SG-Emojis denotes fusing Word2Vec Skip Gram with Emojis Frequencies; whereas CBOW-SG-Emojis denotes fusing Word2Vec CBOW, Word2Vec Skip Gram and Emojis Frequencies). We can observe that the performance has remarkably improved, with a highest accuracy of $85.41\% \pm 2.59$ when using two representations (SG-Emojis) at the score level, using any fusion rule sum, prod or max. We also noticed that although combining the three feature representations (CBOW-SG-Emojis)

Table 4.4: Results using emojis based features

clf	Recall	Prc	F_1	GM	Acc
Feature Extraction Method: Emoji frequency					
SGD	74.13 \pm 3.71	76.10 \pm 3.58	73.28 \pm 4.58	73.02 \pm 5.52	74.13 \pm 3.71
SVC	77.96 \pm 2.12	79.11 \pm 1.81	78.07 \pm 2.12	78.50 \pm 1.95	77.96 \pm 2.12
GNB	57.59 \pm 3.54	71.32 \pm 2.83	54.07 \pm 4.63	62.42 \pm 3.15	57.59 \pm 3.54
NN	71.45 \pm 2.80	71.37 \pm 2.94	71.25 \pm 2.86	70.20 \pm 3.04	71.45 \pm 2.80
LR	78.53 \pm 2.33	79.75 \pm 2.41	78.64 \pm 2.33	79.12 \pm 2.42	78.53 \pm 2.33
DT	75.52 \pm 1.88	76.33 \pm 1.88	75.64 \pm 1.87	75.76 \pm 1.93	75.52 \pm 1.88
RF	76.66 \pm 1.76	77.24 \pm 1.69	76.74 \pm 1.75	76.65 \pm 1.76	76.66 \pm 1.76
GB	75.66 \pm 2.62	76.12 \pm 2.79	74.87 \pm 2.84	73.19 \pm 2.94	75.66 \pm 2.62
VE	77.05 \pm 2.06	78.00 \pm 2.19	76.99 \pm 2.00	76.84 \pm 2.19	77.05 \pm 2.06
SE	76.52 \pm 2.73	77.62 \pm 2.77	76.43 \pm 2.79	76.36 \pm 2.97	76.52 \pm 2.73
Feature Extraction Method: Lexicon-based					
SGD	74.18 \pm 3.50	77.54 \pm 3.25	73.05 \pm 5.02	73.47 \pm 6.12	74.18 \pm 3.50
SVC	77.96 \pm 1.88	79.70 \pm 1.77	78.08 \pm 1.89	78.91 \pm 1.82	77.96 \pm 1.88
GNB	57.69 \pm 3.22	71.21 \pm 2.71	54.28 \pm 4.20	62.47 \pm 2.88	57.69 \pm 3.22
NN	71.21 \pm 2.66	71.16 \pm 2.71	70.88 \pm 2.53	69.65 \pm 2.40	71.21 \pm 2.66
LR	76.33 \pm 1.65	79.61 \pm 1.71	76.38 \pm 1.71	77.99 \pm 1.58	76.33 \pm 1.65
DT	74.95 \pm 2.83	75.07 \pm 2.83	74.94 \pm 2.79	74.32 \pm 2.84	74.95 \pm 2.83
RF	75.61 \pm 2.13	75.67 \pm 2.19	75.54 \pm 2.14	74.80 \pm 2.27	75.61 \pm 2.13
GB	75.46 \pm 2.76	75.95 \pm 2.98	74.65 \pm 2.97	72.96 \pm 3.07	75.46 \pm 2.76
VE	76.81 \pm 2.27	78.54 \pm 2.38	76.81 \pm 2.39	77.39 \pm 2.79	76.81 \pm 2.27
SE	76.76 \pm 3.23	79.33 \pm 2.92	76.71 \pm 3.44	77.77 \pm 3.38	76.76 \pm 3.23
Feature Extraction Method: Emojis Embedding CBOW					
SGD	73.65 \pm 3.49	76.63 \pm 2.55	72.62 \pm 4.83	72.88 \pm 5.36	73.65 \pm 3.49
SVC	78.67 \pm 1.39	79.31 \pm 1.52	78.77 \pm 1.39	78.84 \pm 1.53	78.67 \pm 1.39
GNB	76.81 \pm 2.21	76.87 \pm 2.27	76.75 \pm 2.25	76.05 \pm 2.42	76.81 \pm 2.21
NN	73.46 \pm 3.25	73.57 \pm 3.29	73.32 \pm 3.20	72.44 \pm 3.22	73.46 \pm 3.25
LR	79.10 \pm 2.17	79.77 \pm 2.32	79.20 \pm 2.18	79.29 \pm 2.41	79.10 \pm 2.17
DT	74.66 \pm 2.30	75.15 \pm 2.36	74.75 \pm 2.28	74.54 \pm 2.37	74.66 \pm 2.30
RF	77.96 \pm 2.86	78.18 \pm 3.01	77.98 \pm 2.89	77.60 \pm 3.14	77.96 \pm 2.86
GB	77.72 \pm 2.85	77.82 \pm 2.94	77.69 \pm 2.87	77.10 \pm 3.04	77.72 \pm 2.85
SE	76.62 \pm 2.98	77.21 \pm 2.62	76.32 \pm 3.48	75.70 \pm 3.93	76.62 \pm 2.98
Feature Extraction Method: Emojis Embedding SG					
SGD	73.70 \pm 3.69	77.04 \pm 2.63	72.61 \pm 4.85	72.94 \pm 5.15	73.70 \pm 3.69
SVC	78.62 \pm 1.57	79.24 \pm 1.47	78.72 \pm 1.56	78.77 \pm 1.54	78.62 \pm 1.57
GNB	77.62 \pm 1.73	77.77 \pm 1.80	77.60 \pm 1.76	77.06 \pm 1.94	77.62 \pm 1.73
NN	71.35 \pm 5.02	71.62 \pm 5.59	70.98 \pm 5.21	70.05 \pm 5.71	71.35 \pm 5.02
LR	79.53 \pm 2.03	80.17 \pm 2.16	79.62 \pm 2.03	79.71 \pm 2.22	79.53 \pm 2.03
DT	75.57 \pm 2.54	76.11 \pm 2.74	75.65 \pm 2.55	75.51 \pm 2.80	75.57 \pm 2.54
RF	77.76 \pm 2.73	77.96 \pm 2.90	77.76 \pm 2.78	77.30 \pm 3.06	77.76 \pm 2.73
GB	78.00 \pm 2.09	78.15 \pm 2.24	77.97 \pm 2.12	77.41 \pm 2.34	78.00 \pm 2.09
VE	77.91 \pm 2.76	78.71 \pm 2.16	77.93 \pm 2.77	77.92 \pm 2.42	77.91 \pm 2.76
SE	77.72 \pm 2.71	78.59 \pm 3.08	77.65 \pm 2.85	77.53 \pm 3.15	77.72 \pm 2.71

Table 4.5: Fusion of Word2Vec CBOW, Word2Vec Skip Gram and Emojis frequencies at feature, score and decision levels

Modality	Level/Function	Recall	Prc	F_1	Gm	Acc
CBOW, Emojis	C-E	83.83 \pm 2.64	84.00 \pm 2.78	83.85 \pm 2.65	83.57 \pm 2.90	83.83 \pm 2.64
	SUM(C,E)	84.31 \pm 2.09	84.31 \pm 2.10	84.25 \pm 2.13	83.58 \pm 2.31	84.31 \pm 2.09
	PROD(C,E)	84.31 \pm 2.09	84.31 \pm 2.10	84.25 \pm 2.13	83.58 \pm 2.31	84.31 \pm 2.09
	MAX(C,E)	84.31 \pm 2.09	84.31 \pm 2.10	84.25 \pm 2.13	83.58 \pm 2.31	84.31 \pm 2.09
SG, Emojis	S-E	83.74 \pm 2.78	83.88 \pm 2.79	83.75 \pm 2.77	83.45 \pm 2.87	83.74 \pm 2.78
	SUM(S,E)	85.41 \pm 2.59	85.43 \pm 2.60	85.37 \pm 2.62	84.80 \pm 2.83	85.41 \pm 2.59
	PROD(S,E)	85.41 \pm 2.59	85.43 \pm 2.60	85.37 \pm 2.62	84.80 \pm 2.83	85.41 \pm 2.59
	MAX(S,E)	85.41 \pm 2.59	85.43 \pm 2.60	85.37 \pm 2.62	84.80 \pm 2.83	85.41 \pm 2.59
CBOW, SG, Emojis	C-S-E	83.41 \pm 3.21	83.48 \pm 3.18	83.41 \pm 3.20	83.01 \pm 3.26	83.41 \pm 3.21
	SUM(C,S,E)	84.60 \pm 2.56	84.64 \pm 2.59	84.58 \pm 2.58	84.14 \pm 2.74	84.60 \pm 2.56
	PROD(C,S,E)	84.74 \pm 2.29	84.76 \pm 2.30	84.72 \pm 2.30	84.23 \pm 2.43	84.74 \pm 2.29
	MAX(C,S,E)	85.08 \pm 2.34	85.07 \pm 2.36	85.03 \pm 2.37	84.44 \pm 2.53	85.08 \pm 2.34
	MOD(C,S,E)	83.07 \pm 2.77	83.14 \pm 2.82	83.06 \pm 2.80	82.62 \pm 2.99	83.07 \pm 2.77

has slightly lower accuracy than combining only Skip Gram and Emojis.

Figure 4.8 compares the performance in terms of ROC curves and the AUCs for five sentiment classification models using SVM . The feature extractors covers three baseline methods each using a single modality: CBOW, Skip Gram (SG), and Emojis. It also shows the best two score-level fusion methods using two modalities: CBOW with Emojis (C-E) and Skip Gram with Emojis (S-E).

4.3.2 Handling Emojis Class Imbalance Issue

The first observation is that users tend to use emojis when they are happy to express their positive opinions. This observation confirms the findings reported in the study of [195]. As a result, this part presents a method to address it as a class imbalance problem. The proposed method is based on bagging algorithm and oversampling methods to build multiple models from the training dataset. The layout of the proposed method is depicted in Figure 4.9. It compares three approaches based on: (a) single classifiers, (b) Conventional bagging classifiers, and (c) balanced bagging classifiers.

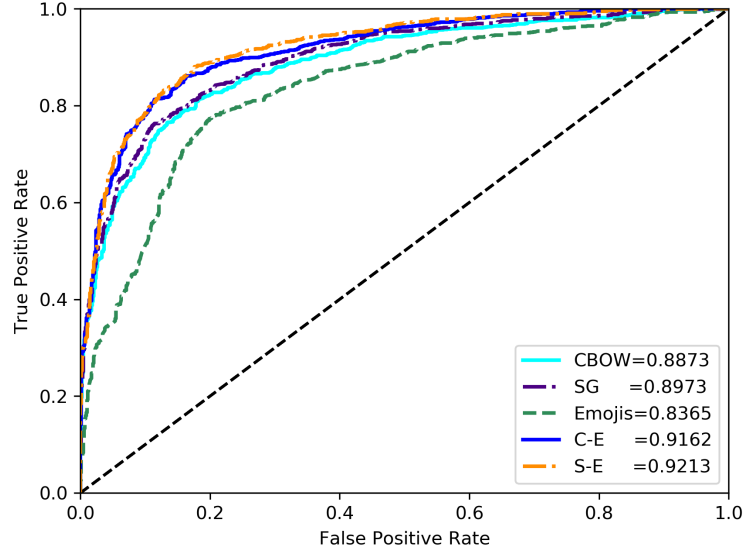


Figure 4.8: Performance comparison using ROC curves and AUCs for five sentiment classification models using SVM

The single classifiers are used as base learners for each subset in the conventional and balanced bagging based ensembles.

- **Single classifiers:** Three popular machine-learning classifiers are considered: k -NN, GNB and DT. They are first used as single classifiers to have a baseline for comparison. Implementing this task results in generating three different models to solve the same problem, using the single classifiers.
- **Conventional bagging classifier:** A conventional bagging-based ensemble approach generates different randomly selected subsets of data then builds several estimators on each subset. Three different conventional bagging methods are considered; each of which has the same structurer but different base classifier. In other words, one bagging classifier uses NB as a base estimator, a second bagging classifier uses k -NN as a base estimator, and a third bagging classifier

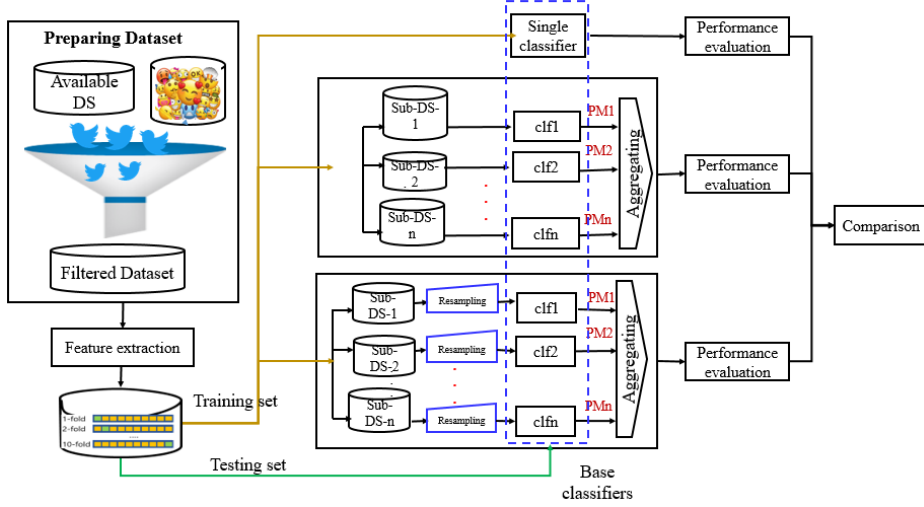


Figure 4.9: Layout of the proposed approach for handling emojis imbalance issue

uses DT as a base estimator. This gives us another baseline for comparison.

- **Balanced bagging classifier:** In the case of imbalanced datasets, a conventional bagging method is not allowed to balance each subset of data since it results in favoring the majority class. To avoid this problem, each subset of the dataset needs to be balanced before training each estimator. To do so, oversampling methods are applied on each training subset. Since there are three different base classifiers, different models are generated to solve the same problem. The first model is a balanced bagging classifier using NB as base estimators, the second model is a balanced bagging classifier using k -NN as base estimators, and the third model is a balanced bagging classifier using decision trees as base estimators.

For k -NN, $k = 1$ is set in our experiments. Additionally, for decision tree, CART algorithm is used with a Gini index. The minimum sample split is set to two. Twenty

Table 4.6: Summary of the used datasets with different imbalance ratio

Dataset	Negative	Positive	Total	IR
Dataset I	377	724	1101	1.92
Dataset II	242	724	966	2.99
Dataset III	100	724	824	7.24
Dataset VI	50	724	774	14.48

estimators are considered for each bagging classifier. In addition to the original data (Dataset I), three other datasets are created with higher imbalance ratios, denoted as Dataset II, Dataset III and Dataset VI. They were created through eliminating instances randomly from the minority class (negative) in Dataset I. The imbalance ratio ranges from almost two to over 14 and the distributions of all datasets are shown in Table 4.6. In the case of a balanced bagging classifier, oversampling methods are only applied to the training subsets since it is unreasonable to validate models using synthetic instances.

Twenty-four models are generated for each dataset. The results are shown in Tables 4.7, 4.8 and 4.9 for Dataset I, Dataset II, Dataset III and Dataset VI, respectively. For each table, the results are represented in three parts. Each part represents a single classifier, conventional bagging classifier with using the same single classifier as estimator and the balanced bagging classifier with different oversampling methods. For each measure the highest, results are represented in bold and the highest results per each dataset are denoted with a star “*” sign. It is clear that the highest results are obtained using the bagging classifier with oversampling methods. For Dataset I, the highest results are obtained using the balanced bagging classifier with SMOTE-SVM in case of using decision tree as base classifier. For Dataset II, the highest results are obtained using the balanced bagging classifier with SMOTE-B1 and SMOTE-B2 in

Table 4.7: Results using the original dataset (Dataset I)

Method	F1	MCC	GM	AUC
GNB				
GNB	49.16 \pm 5.09	27.68 \pm 8.19	61.23 \pm 4.04	78.08 \pm 5.07
Bagging	50.64 \pm 4.33	29.17 \pm 7.67	62.12 \pm 3.69	75.34 \pm 6.11
Bagging-ROS	51.31 \pm 5.03	29.10 \pm 7.78	62.31 \pm 3.91	75.06 \pm 6.26
Bagging-SMOTE	52.50 \pm 4.70	28.97 \pm 8.18	62.63 \pm 4.02	74.63 \pm 6.01
Bagging-SMOTE-B1	71.46 \pm 3.23	41.33 \pm 6.88	71.67 \pm 3.62	78.41 \pm 3.70
Bagging-SMOTE-B2	52.63 \pm 5.60	27.40 \pm 9.01	62.16 \pm 4.60	74.70 \pm 5.99
Bagging-SMOTE-SVM	52.13 \pm 5.07	28.57 \pm 8.15	62.40 \pm 4.14	75.10 \pm 5.44
Bagging-ADASYN	49.03 \pm 5.16	27.06 \pm 9.33	60.98 \pm 4.41	74.45 \pm 6.01
k-NN				
kNN	69.76 \pm 3.61	33.08 \pm 8.12	66.11 \pm 4.23	66.24 \pm 4.18
Bagging	73.35 \pm 4.89	40.80 \pm 10.92	70.24 \pm 5.49	76.98 \pm 4.39
Bagging-ROS	69.04 \pm 5.20	31.61 \pm 11.71	62.91 \pm 5.45	75.99 \pm 4.97
Bagging-SMOTE	73.28 \pm 4.39	42.11 \pm 9.13	71.53 \pm 4.69	78.23 \pm 4.10
Bagging-SMOTE-B1	71.46 \pm 3.23	41.33 \pm 6.88	71.67 \pm 3.62	78.41 \pm 3.70
Bagging-SMOTE-B2	73.25 \pm 4.24	41.35 \pm 9.73	71.02 \pm 5.16	77.26 \pm 3.99
Bagging-SMOTE-SVM	74.13 \pm 3.55	43.62 \pm 7.70	72.22 \pm 3.99	77.97 \pm 4.16
Bagging-ADASYN	72.75 \pm 5.29	41.42 \pm 10.87	71.32 \pm 5.57	77.19 \pm 4.68
DT				
DT	74.22 \pm 5.14	44.13 \pm 11.17	72.48 \pm 5.73	75.40 \pm 6.12
Bagging	74.37 \pm 4.04	43.37 \pm 8.97	71.44 \pm 4.71	78.44 \pm 4.80
Bagging-ROS	74.18 \pm 3.35	44.60 \pm 7.17	73.00 \pm 3.71	78.71 \pm 3.54
Bagging-SMOTE	74.36 \pm 3.25	44.44 \pm 6.97	72.74 \pm 3.60	79.04* \pm 4.13
Bagging-SMOTE-B1	71.46 \pm 3.23	41.33 \pm 6.88	71.67 \pm 3.62	78.41 \pm 3.70
Bagging-SMOTE-B2	74.51 \pm 3.44	44.58 \pm 7.47	72.76 \pm 3.83	78.63 \pm 3.71
Bagging-SMOTE-SVM	74.63* \pm 2.91	45.08* \pm 6.17	73.08* \pm 3.19	79.00* \pm 4.04
Bagging-ADASYN	71.76 \pm 3.69	42.96 \pm 7.58	72.57 \pm 3.98	78.37 \pm 3.69

case of using decision tree as base classifier. For Dataset III, the highest results are obtained using the balanced bagging classifier with SMOTE-ROS (in terms of AUC) and Bagging-ADASYN (in terms of MCC and GM) in case of using decision tree as base classifier.

4.4 Summary

In summary, the experimental results illustrate:

- Emojis are capable to detect polarity in microblogs and outperform *tf-idf* and LSA textual features and structural features significantly.

Table 4.8: Results using highly imbalanced dataset (Dataset II)

Method	F1	MCC	GM	AUC
GNB				
GNB	43.11 \pm 4.08	17.16 \pm 6.55	56.46 \pm 3.26	74.68 \pm 4.44
Bagging	46.38 \pm 3.23	20.34 \pm 6.38	58.57 \pm 3.04	73.35 \pm 4.95
Bagging-ROS	47.09 \pm 3.15	20.55 \pm 7.00	58.86 \pm 3.37	73.57 \pm 4.77
Bagging-SMOTE	50.23 \pm 2.73	22.28 \pm 6.86	60.41 \pm 3.32	73.23 \pm 4.71
Bagging-SMOTE-B1	73.61 \pm 3.58	37.78* \pm 10.55	70.86* \pm 6.17	75.96 \pm 5.17
Bagging-SMOTE-B2	47.20 \pm 3.59	19.89 \pm 7.95	58.65 \pm 3.89	70.93 \pm 4.60
Bagging-SMOTE-SVM	51.94 \pm 3.82	24.18 \pm 6.66	61.66 \pm 3.48	74.02 \pm 4.85
Bagging-ADASYN	44.32 \pm 3.86	18.21 \pm 7.09	57.18 \pm 3.42	72.48 \pm 5.98
k-NN				
k-NN	72.23 \pm 4.53	25.10 \pm 12.62	60.76 \pm 6.73	61.89 \pm 6.20
Bagging	74.84 \pm 2.95	32.61 \pm 8.11	64.87 \pm 4.86	74.76 \pm 4.09
Bagging-ROS	68.31 \pm 3.68	12.34 \pm 12.04	50.57 \pm 5.80	70.98 \pm 5.49
Bagging-SMOTE	74.90 \pm 2.90	34.44 \pm 8.20	66.84 \pm 5.18	75.66 \pm 4.47
Bagging-SMOTE-B1	73.61 \pm 3.58	37.78* \pm 10.55	70.86* \pm 6.17	75.96 \pm 5.17
Bagging-SMOTE-B2	75.16 \pm 2.88	33.95 \pm 8.33	66.13 \pm 4.95	75.06 \pm 4.59
Bagging-SMOTE-SVM	75.13 \pm 2.80	34.47 \pm 7.77	66.89 \pm 4.58	75.96 \pm 4.01
Bagging-ADASYN	74.35 \pm 3.30	32.83 \pm 8.92	66.15 \pm 5.34	75.60 \pm 3.80
DT				
DT	73.89 \pm 3.40	29.47 \pm 9.26	63.11 \pm 5.09	70.96 \pm 6.49
Bagging	74.08 \pm 4.08	29.51 \pm 10.93	62.72 \pm 5.63	72.92 \pm 7.09
Bagging-ROS	72.96 \pm 4.34	34.94 \pm 10.94	69.10 \pm 6.13	75.96 \pm 5.27
Bagging-SMOTE	74.02 \pm 4.40	35.17 \pm 9.96	68.50 \pm 5.14	75.77 \pm 5.90
Bagging-SMOTE-B1	73.61 \pm 3.58	37.78* \pm 10.55	70.86* \pm 6.17	75.96 \pm 5.17
Bagging-SMOTE-B2	75.32* \pm 3.33	36.31 \pm 9.67	68.43 \pm 5.69	76.51* \pm 4.54
Bagging-SMOTE-SVM	74.64 \pm 4.14	36.01 \pm 10.08	68.70 \pm 5.47	75.55 \pm 5.66
Bagging-ADASYN	71.47 \pm 4.19	35.74 \pm 10.61	70.03 \pm 6.12	76.35 \pm 4.48

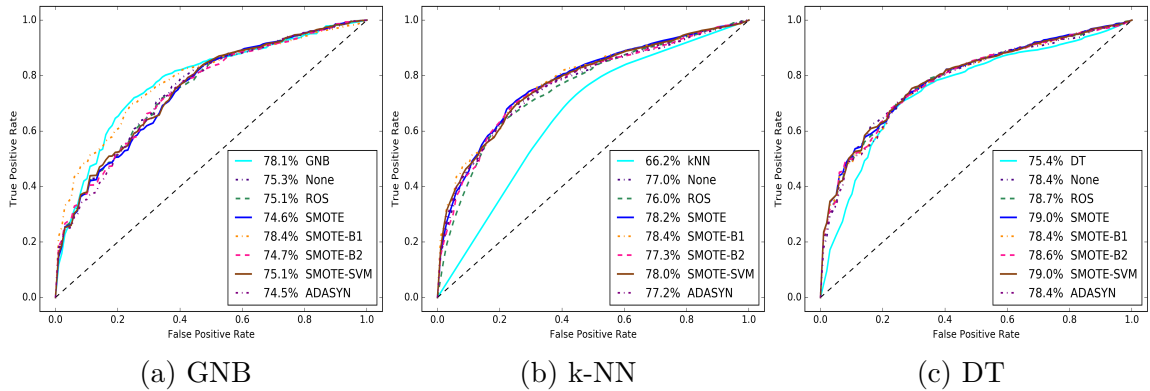


Figure 4.10: ROC and AUC for Dataset I for the single classifier, conventional bagging classifier and balanced bagging classifier

Table 4.9: Results using the more highly imbalanced dataset (Dataset III)

Method	F1	MCC	GM	AUC
GNB				
GNB	67.35 \pm 2.44	25.52 \pm 5.28	68.76 \pm 3.57	76.49 \pm 4.43
Bagging	68.91 \pm 2.80	26.97 \pm 5.32	69.93 \pm 3.72	76.89 \pm 3.00
Bagging-ROS	69.02 \pm 2.66	27.06 \pm 5.26	70.01 \pm 3.66	76.73 \pm 2.86
Bagging-SMOTE	69.64 \pm 2.51	27.64 \pm 5.12	70.46 \pm 3.57	76.87 \pm 2.54
Bagging-SMOTE-B1	82.02 \pm 2.70	23.04 \pm 14.99	60.14 \pm 12.66	79.18 \pm 3.77
Bagging-SMOTE-B2	49.97 \pm 4.13	13.94 \pm 6.57	57.50 \pm 4.30	69.22 \pm 6.48
Bagging-SMOTE-SVM	69.74 \pm 2.49	27.74 \pm 5.20	70.54 \pm 3.64	76.86 \pm 2.87
Bagging-ADASYN	50.68 \pm 2.88	14.39 \pm 6.20	57.96 \pm 3.81	71.44 \pm 4.07
k-NN				
kNN	82.66 \pm 2.98	26.67 \pm 12.73	62.64 \pm 9.94	65.30 \pm 7.92
Bagging	83.34 \pm 2.70	19.03 \pm 14.95	52.55 \pm 10.99	77.18 \pm 4.20
Bagging-ROS	84.41* \pm 2.81	19.89 \pm 16.77	48.71 \pm 10.03	77.12 \pm 5.42
Bagging-SMOTE	84.02 \pm 3.73	25.77 \pm 17.37	58.34 \pm 12.06	79.65 \pm 3.91
Bagging-SMOTE-B1	82.02 \pm 2.70	23.04 \pm 14.99	60.14 \pm 12.66	79.18 \pm 3.77
Bagging-SMOTE-B2	82.98 \pm 3.44	23.08 \pm 16.75	58.17 \pm 12.49	77.81 \pm 3.65
Bagging-SMOTE-SVM	83.30 \pm 3.78	22.22 \pm 17.72	56.10 \pm 12.42	78.51 \pm 2.94
Bagging-ADASYN	83.94 \pm 2.89	25.63 \pm 14.67	59.07 \pm 10.51	78.48 \pm 3.95
DT				
DT	83.73 \pm 2.06	23.25 \pm 11.30	56.76 \pm 8.75	74.83 \pm 5.82
Bagging	83.41 \pm 2.62	18.33 \pm 13.97	50.17 \pm 9.53	77.64 \pm 4.92
Bagging-ROS	80.06 \pm 3.24	27.17 \pm 11.64	66.52 \pm 8.61	79.85* \pm 3.27
Bagging-SMOTE	81.07 \pm 2.95	18.97 \pm 15.54	57.22 \pm 13.08	77.66 \pm 3.75
Bagging-SMOTE-B1	82.02 \pm 2.70	23.04 \pm 14.99	60.14 \pm 12.66	79.18 \pm 3.77
Bagging-SMOTE-B2	82.91 \pm 3.03	22.49 \pm 14.11	57.61 \pm 11.18	79.59 \pm 3.89
Bagging-SMOTE-SVM	81.84 \pm 3.14	18.76 \pm 15.09	55.56 \pm 12.44	77.70 \pm 4.32
Bagging-ADASYN	78.94 \pm 2.38	32.31* \pm 7.10	72.36* \pm 4.96	79.06 \pm 2.44

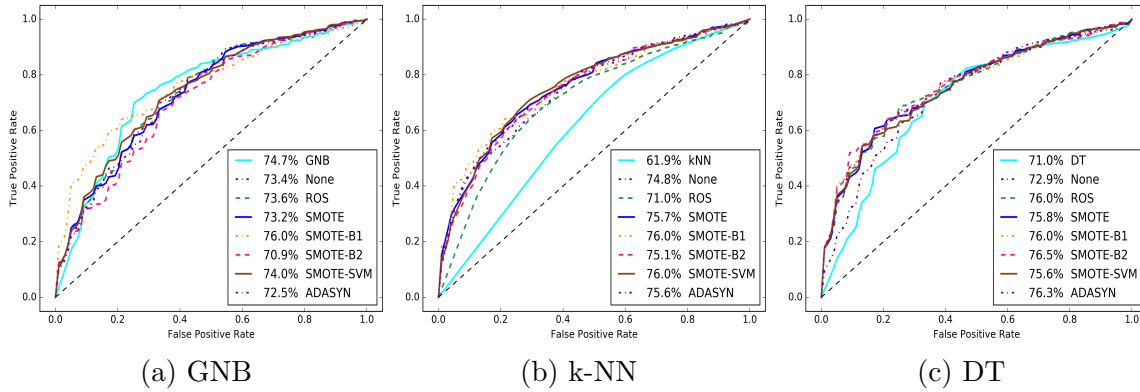


Figure 4.11: ROC and AUC for Dataset II for the single classifier, conventional bagging classifier and balanced bagging classifier

Table 4.10: Results using the more highly imbalanced dataset (Dataset IV)

Method	F1	MCC	GM	AUC
GNB				
GNB	72.73 \pm 5.41	14.35 \pm 7.74	63.35 \pm 7.56	68.92 \pm 9.83
Bagging	74.31 \pm 5.06	14.65 \pm 7.59	63.32 \pm 7.68	70.65 \pm 10.40
Bagging-ROS	74.31 \pm 5.06	14.65 \pm 7.59	63.32 \pm 7.68	71.36 \pm 10.19
Bagging-SMOTE	74.70 \pm 5.04	14.98 \pm 7.53	63.58 \pm 7.68	70.62 \pm 9.84
Bagging-SMOTE-B1	88.14 \pm 2.92	7.30 \pm 15.36	37.88 \pm 16.10	74.05 \pm 10.00
Bagging-SMOTE-B2	54.70 \pm 5.33	3.78 \pm 8.70	52.39 \pm 8.11	59.93 \pm 12.86
Bagging-SMOTE-SVM	75.11 \pm 4.87	14.39 \pm 7.75	63.01 \pm 7.80	70.75 \pm 9.74
Bagging-ADASYN	55.96 \pm 5.02	5.36 \pm 10.36	53.96 \pm 9.56	64.07 \pm 12.96
k-NN				
kNN	87.38 \pm 2.33	13.90 \pm 11.54	50.97 \pm 12.10	58.75 \pm 6.80
Bagging	90.12* \pm 1.01	6.48 \pm 11.07	33.62 \pm 11.41	76.40 \pm 7.78
Bagging-ROS	89.96 \pm 0.62	3.54 \pm 8.71	31.24 \pm 10.62	72.88 \pm 9.53
Bagging-SMOTE	89.43 \pm 1.50	10.96 \pm 16.77	42.04 \pm 17.45	76.89* \pm 7.01
Bagging-SMOTE-B1	88.14 \pm 2.92	7.30 \pm 15.36	37.88 \pm 16.10	74.05 \pm 10.00
Bagging-SMOTE-B2	89.31 \pm 1.61	12.07 \pm 17.49	43.46 \pm 18.53	75.08 \pm 7.17
Bagging-SMOTE-SVM	89.27 \pm 1.51	7.24 \pm 17.86	37.36 \pm 18.60	76.48 \pm 8.85
Bagging-ADASYN	90.06 \pm 1.65	14.97 \pm 16.65	44.66 \pm 16.45	75.14 \pm 9.37
DT				
DT	89.61 \pm 1.36	8.96 \pm 18.26	39.02 \pm 19.77	70.47 \pm 13.81
Bagging	90.00 \pm 0.87	3.98 \pm 10.43	31.25 \pm 10.79	74.42 \pm 10.46
Bagging-ROS	87.30 \pm 2.62	24.53* \pm 10.65	64.51* \pm 10.95	74.86 \pm 10.75
Bagging-SMOTE	87.95 \pm 2.55	2.28 \pm 12.50	33.17 \pm 15.83	73.55 \pm 10.26
Bagging-SMOTE-B1	88.14 \pm 2.92	7.30 \pm 15.36	37.88 \pm 16.10	74.05 \pm 10.00
Bagging-SMOTE-B2	88.87 \pm 2.07	3.89 \pm 13.90	33.65 \pm 16.01	73.98 \pm 9.98
Bagging-SMOTE-SVM	88.80 \pm 2.52	4.09 \pm 14.28	33.62 \pm 16.10	74.03 \pm 10.05
Bagging-ADASYN	85.86 \pm 3.37	20.95 \pm 13.42	62.41 \pm 12.46	73.52 \pm 10.19

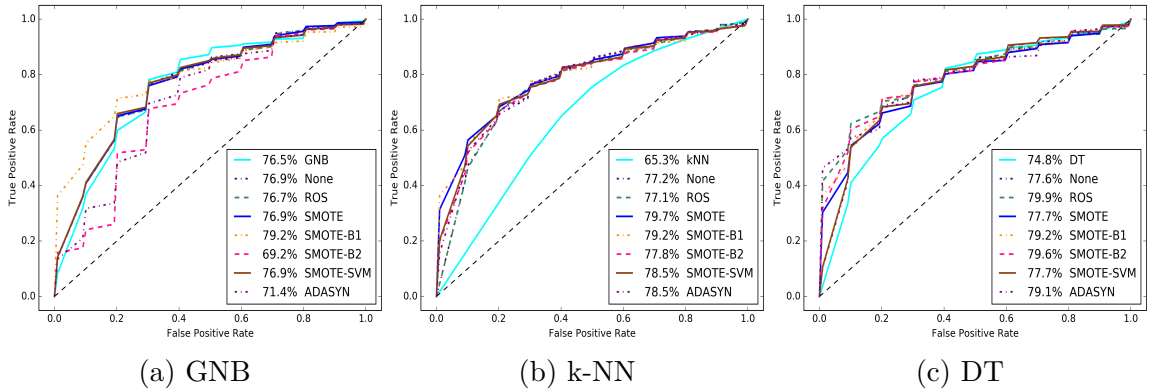


Figure 4.12: ROC and AUC for Dataset III for the single classifier, conventional bagging classifier and balanced bagging classifier

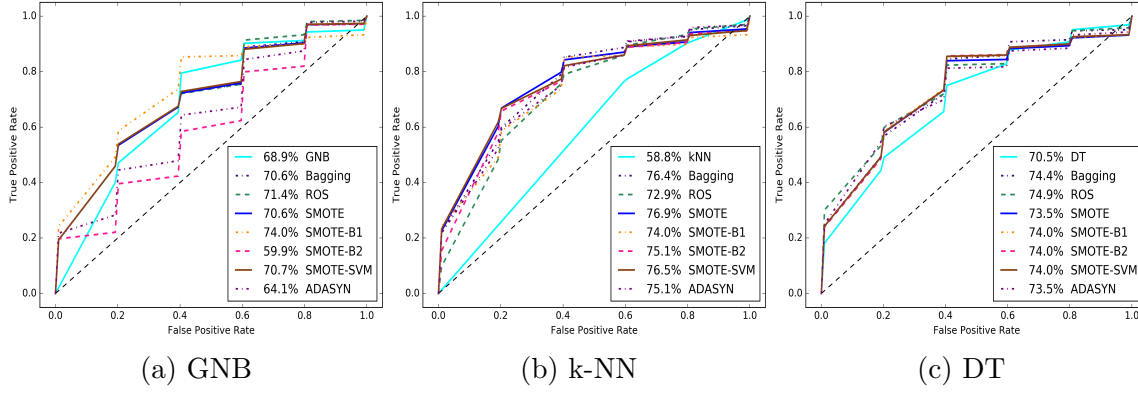


Figure 4.13: ROC and AUC for Dataset IV for the single classifier, conventional bagging classifier and balanced bagging classifier

- Emojis-based features generated using skip-gram outperform other types of emojis-based features.
- Word embedding based textual features achieve the highest results in case of unimodal approaches.
- Fusing emojis with text leads to improve the performances where the highest results are attained using score level fusion.
- The proposed approach to handle emojis imbalance class problem performs better than baselines in most of the considered cases.

CHAPTER 5

MULTIMODAL SENTIMENT ANALYSIS

Most of the existing techniques in the literature for sentiment analysis have focused on text modality. Recently, some researchers have been motivated to other modalities such as audio and visual but the work is still in its early stages. This chapter presents a multimodal Arabic sentiment analysis approach of audio, textual and visual. It also presents a new multimodal sentiment analysis dataset of Arabic video opinions and investigates different features and explores different fusion techniques with intensive empirical analysis.

5.1 Multimodal Sentiment Analysis Framework

The proposed framework is composed of several modules illustrated in Figure 5.1:

- Data acquisition and preparation module: This module contains several tasks including video collection, segmentation, video and audio separation, transcrip-

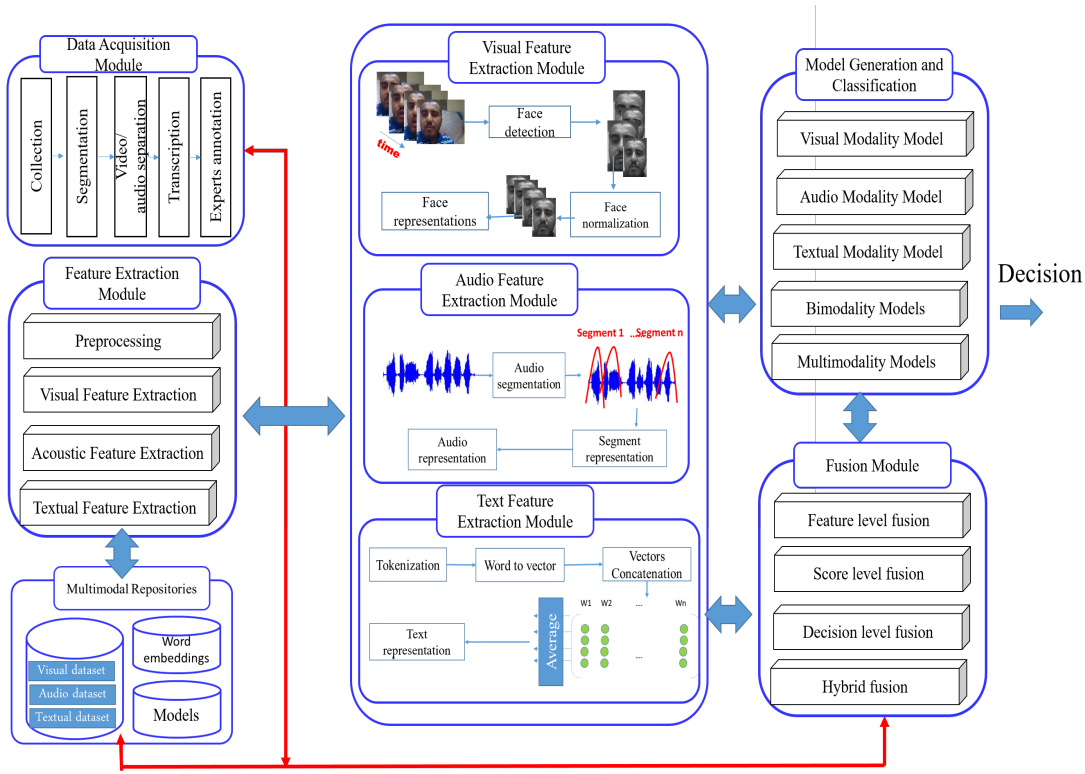


Figure 5.1: Multimodal Arabic opinion mining framework

tion and annotation.

- **Feature extraction module:** Feature extraction is a significant task in machine learning approaches by which each instance input is mapped into a representation of its characteristics. This module includes several sub-modules for preprocessing, visual feature extraction, audio feature extraction, textual feature extraction and fusion. Some preprocessing steps are conducted. For visual modality, face region is first detected then RGB color image is converted to gray scale. Each audio input is in ‘WAV’ format of 256 bit, 48000 Hz sampling frequency and a mono channel. Preprocessing operations including normalizing Alefs and Tah Marbotah are carried out on the texts. A feature extractor is

constructed for each modality. The audio feature extractor constructs feature vectors of 68 features for each instance. Moreover, a textual feature extractor is implemented to extract textual features based on word embeddings, as described in Chapter 3. Optical flow features are considered to represent the visual modality. Different parameters are investigated to generate the visual feature vectors.

- Fusion module: Different fusion levels are investigated and evaluated including: feature level, score level and decision level. In addition, multi-level fusion is proposed to fuse different modalities. To our knowledge, this is the first study to conduct extensive evaluation and exploration for such fusion methods in sentiment analysis.
- Model generation and classification module: The generated feature vectors individually and in combinations are used to train machine learning classifiers that can detect sentiments, from audio, textual and visual modalities. Several evaluation metrics are considered to compare various models.

The proposed methodology is designed to answer the following research questions:

RQ1: Given textual, audio and visual modalities extracted from Arabic opinion videos, what is the most accurate modality to detect the speakers' sentiment?

RQ2: What is the effect of combining different modalities to detect sentiments from Arabic opinion videos?

Table 5.1: SADAM dataset statistics

Sentiment		Gender		Age-group				Dialect			
		Female	Male	AGA	AGB	AGC	AGD	Egyptian	Gulf	Levantine	Maghrebi
Positive	274	134	140	103	113	47	11	106	106	33	16
Negative	250	82	168	25	46	95	84	143	61	46	13
Total	524	216	308	128	159	142	95	249	167	79	29

Table 5.2: SADAM dataset description

Statistical measure	Value
Total no. videos	63
Total no. videos expressed by male	37
Total no. videos expressed by female	26
Total no. distinct speakers	59
Total no. opinion segments	524
Total no. positive segments	274
Total no. negative segments	250
Average no. opinion segments in video	8.32
Average length of opinion segments (seconds)	5.29
Average no. of video frames	137.24
Average word count per opinion segments	12.52
Total no. words in segments	6562
Total no. unique words in opinion segments	2774
Total no. words appears in segments at least 5 times in the dataset	491

RQ3: What is most efficient fusion method to combine the aforementioned modalities?

RQ4: Does multi-level hybrid fusion method improve the results comparing to the single-level fusion methods (feature fusion, decision fusion and score fusion)?

RQ5: What is the impact of demographic segmentation on sentiment analysis? (this will be covered in Chapter 6)

5.2 Dataset Preparation and Collection

The effectiveness of a sentiment analysis system relies on the data collection methodology. There are two main methodologies for constructing multimodal sentiment analysis datasets. Data might be recorded in a controlled or acted environment under special settings, e.g. [116, 118, 130], or videos are recorded in real environments, e.g. [94]. Although the former approach can be more accurate, some subjects may poorly act leading to corrupted training information [13]. Moreover, the latter approach can handle more emotional variability but might suffer from surrounding noise and subsequently is more challenging.

Due to the scarcity of available datasets for multimodal Arabic sentiment analysis, we have built our own dataset from a collection of relevant opinion videos from YouTube. We will refer to this dataset as Sentiment Analysis Dataset for Arabic Multimodal (SADAM), which has been prepared following a similar methodology to [94]. A summary of this dataset is shown in Tables 5.1 and 5.2.

In the following we describe the details for collecting and preparing the dataset. A main goal is to construct as general as possible dataset for Arabic multimodal sentiment analysis in terms of speakers' ages, genders, nationalities, expressed dialects, recording environments, recording devices and expressed topics. The collected videos are expressed by 37 males and 26 females. The speakers are from different Arab countries and of different ages ranged from 15 to 65 years old, approximately. The topics belong to different domains including opinions on products, movies, persons, politics, and cultural views. The contents are expressed in various dialectics. Different

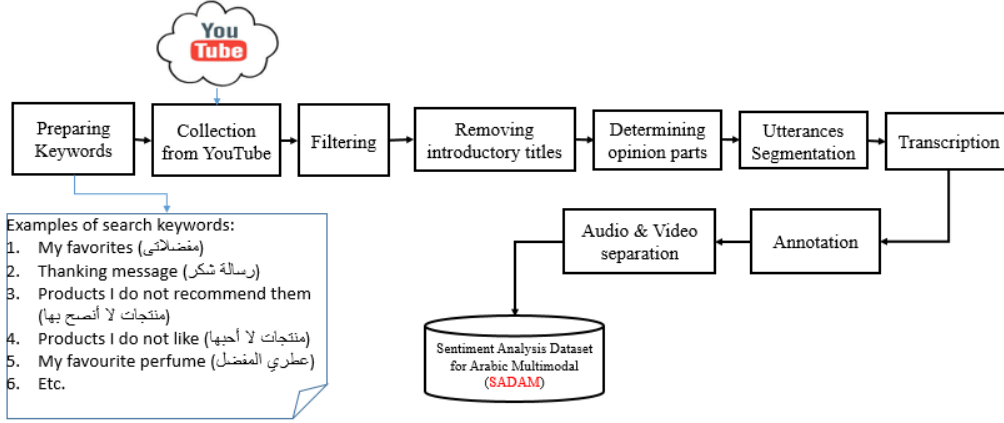


Figure 5.2: Multimodal Arabic sentiment analysis dataset creation process

settings were considered for recorded videos. The collected videos were recorded by users in real environment including houses, studios, offices, cars or outdoors. Users expressed their opinions in different periods. Some videos include more body parts than others. Additionally, some videos have different objects in the background or in the speakers' hands such as reviewed products.

Figure 5.2 shows the main steps of the data collection and preparation, which are summarized as follows. We first prepared a set of search keywords including sentimental phrase such as “my favorites”, “ thanking message”, “my favorite club”, “products I don’t like”, “unrecommended products”, etc. Many videos are found from YouTube. Some of them have been excluded by applying some filtering criteria: videos that don’t present the speaker, videos that include several persons, videos that don’t present nearly all speaker face, etc. Ultimately, we ended up with 63 videos. These videos are processed to get segments that contain single opinions by removing introductory titles and advertisements. Then, each segment is divided into utterances. As a results, we obtained 524 utterances, which are manually converted into transcripts

Table 5.3: Description of the considered speakers' age-groups

Class	Age group	Age interval	Num of samples
AGA: Age-group A	Young adults	15-29 years old	128
AGB: Age-group B	Middle-aged I	30-39 years old	159
AGC: Age-group C	Middle-aged II	40-49 years old	142
AGD: Age-group D	Senior	greater than 49 years old	95
Total			524

(since unlike English these are not available for the Arabic language in YouTube).

Annotation is an important task for constructing labeled datasets for classification problems. This task is conducted carefully and systematically to label our segmented dataset in terms of sentiment and demographic characteristics (gender, age-groups and dialects; to be used later in Chapter 6). For sentiment polarity, two annotators were asked to label each segment as positive or negative based on their perception of the speakers' opinions. A third annotator is involved to break the tie in case of the disagreement between the two annotators.

This process resulted in 274 positive and 250 negative utterances. Gender annotation task is straightforward and the instances are distributed as 308 utterances by males and 216 by females. We adopted four age-groups as described in Table 5.3. For well-known speakers, we looked for their ages in their profiles and assigned their ages by subtracting date of recording videos from their birthdays. For other speakers, four annotators were involved to assign their ages. Following a similar breakdown to [217, 218], we also annotated the dataset into four dialects (Egyptian, Levantine, Gulf and Maghrebi) according to the speakers' nationalities, collected from their profiles, and the annotators' judgments.

5.3 Feature Extraction

In this section, we describe the feature engineering process for audio, textual and visual modalities, respectively.

5.3.1 Acoustic Features

The speech signal features contains most of the emotion specific information and they are classified as prosodic features and spectral features [219]. Prosodic features are influenced by vocal fold activity and appear when we put sounds together in a connected speech such as pitch, ZCR (Zero Crossing Rate), intensity and speech rate [220]. On the other hand, spectral features are influenced by vocal tract activity and are extracted from spectral content of the speech signal, e.g. MFCC (Mel Frequency Cepstral Coefficients), LPCC (Linear Prediction Cepstral Coefficients), LFPC (Log Frequency Power Coefficients) and formants etc. [220].

The input audio is split into frames with size of 50 millisecond with a frame step of 20 millisecond. For each generated frame a set of 34 features are computed, including: (1) ZCR, (2) Energy, (3) Entropy of Energy, (4) Spectral Centroid, (5) Spectral Spread, (6) Spectral Entropy, (7) Spectral Flux, (8) Spectral Rolloff, (9-21) MFCCs, (22-33) Chroma Vector, and (34) Chroma Deviation. Then statistics are computed from each audio's frames to represent the whole audio using one descriptor such as the mean and standard division in our study. Thus, each input audio is represented by 68 (34×2) features (See Figure 5.3). To our knowledge, this is the first time to evaluate this set of acoustic features for multimodal sentiment analysis.

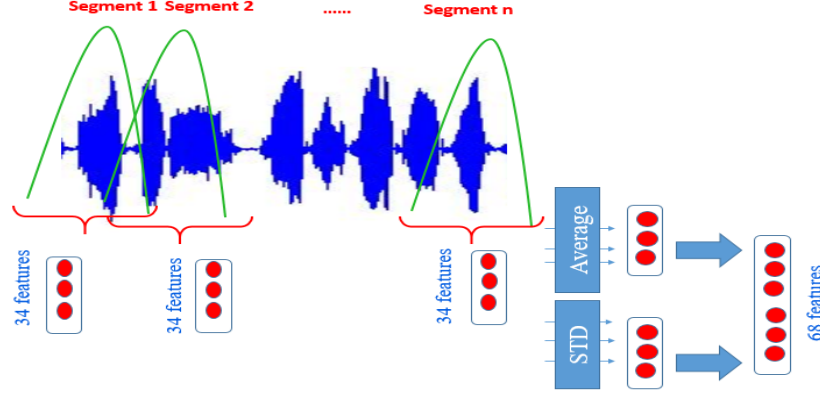


Figure 5.3: Acoustic features extraction process

5.3.2 Transcribed Textual Features

The skip-gram word embedding models [216] are adopted to extract the transcribed textual features. As described in Chapter 4, the models were learnt using a corpus of more than 77,600,000 Arabic tweets posted between 2008 and 2016. They were written in MSA and different Arabic dialects. Different dimensions of 100, 200 and 300 are evaluated. Skip-gram model of 300 dimensionality is selected as the best models in the subsequent experiments. Thus, each instance is represented by a feature vector with size of 300 attributes. This is carried out by calculating the average of the embedding vectors of that sample as described in Section 3.2.3.

5.3.3 Visual Features

Optical flow is a useful technique to represent patterns of apparent motion of objects between adjacent frames in the video [221] and it can be helpful to represent emotional visual patterns. It was reported that the first use of optical flow to track action is attributed to Mase and Pentland [222] in simple manner and static formulation with-

out a physical model. This method involved evaluating the magnitude and direction of motion [223] and representing by a two-dimensional (2D) vector to reflect points movement through two consecutive frames.

The traditional optical flow approach assumes that the pixel intensities of an object do not change between neighboring frames. Thus, a pixel (x, y) at time t will be shifted by dx and dy after dt resulting in same intensity, i.e.

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \quad (5.1)$$

By taking Taylor series approximation of right-hand side and simplifying the equation, we get:

$$I_x u + I_y v + I_t = 0; \quad (5.2)$$

where: $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$, $u = \frac{dx}{dt}$, and $v = \frac{dy}{dt}$

This equation has two unknowns u and v and cannot be solved without additional assumptions. Different methods have been proposed to address this issue and in this thesis we fundamentally follow Lucas-Kanade method [224] which assumes small and approximately constant displacement of image content between consecutive frames (brightness constancy assumption). Thus, the optical flow equation holds true for pixels within a small window (patch) centered at (x, y) and can be solved by least-square fit approach to obtain:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_{x_i}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i f_{y_i}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix} \quad (5.3)$$

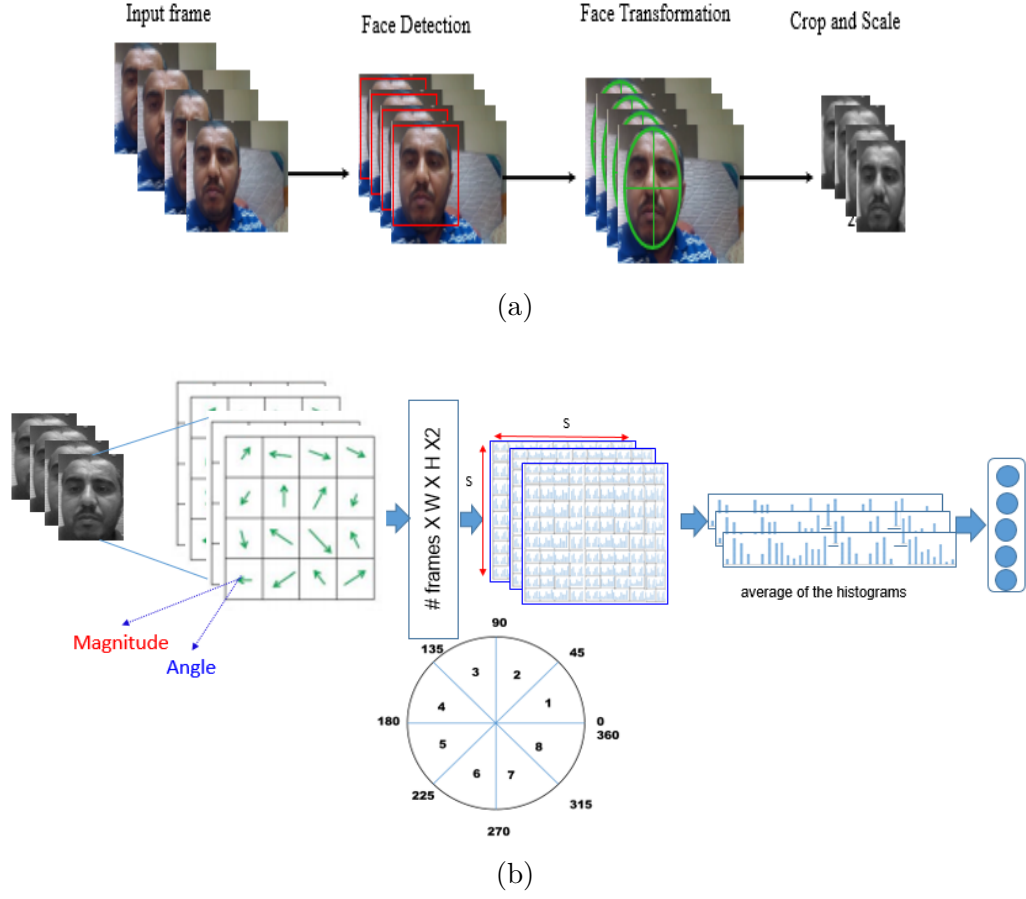


Figure 5.4: (a): Face detection phase, (b) Histogram of optical flow features extraction

However, to go around the small motion assumption, the method is applied iteratively in hierarchical manner yielding what is known as pyramidal Lucas-Kanade method. Another issue is the selection of sparse method to track few control points, such corner points [225], or dense method [226]. In our study, we adopted the later since it is more accurate but slower.

Face detection

The focus in this study is to detect emotion from face only while excluding other body parts. Towards this end, the general frontal face and eye detectors [227] are utilized. The frontal face detector is based on object detection using HAAR feature-

based cascade classifiers [228]. In addition, eye detector detects eye positions which provide significant and useful values to crop and scale the frontal face to a size of considered resolution of 240×320 pixels. This step is considered as a preprocessing step and run once on all data to reduce computation time of the feature extraction.

Features extraction

After detecting the whole face, it is possible to compute the optical flow to capture the evolution of complex motion patterns for the classification of facial expressions. Optical flow is considered to extract the visual features from the videos processed in the previous step. As a result, each point in the frame is represented by two values: magnitude and angle, which describe the vector representing the motion between two consecutive frames. This leads to a huge descriptor of size $NoF \times W \times H \times 2$ to represent each video, where NoF refers to the number of frames in the video and $W \times H$ is the resolution. This large size of dimensionality affects the performance and leads in curse of dimensionality issue. Thus, we need to summarize the generated descriptor as a reduced feature vector to generate machine learning models. Several statistical methods can be used such as average, standard deviation, min, max, etc. In some earlier studies, Histogram of Optical Flow (HOF) has been applied for detecting interaction-level human activities from first-person videos [229] and for determining the movement direction of the object [230] and reported promising results. Similarly, histogram is considered in this study to summarize the high dimensionality descriptor as a single feature vector. Figure. 5.4 depicts the process of visual feature extraction using HOF technique.

Table 5.4: Consumed time for visual feature extraction for all videos

Grid	Directions	Time (Seconds)	No. Features
8×8	6	3888.811	384
	8	5069.572	512
	12	5478.035	768
10×10	6	4130.476	600
	8	6365.948	800
	12	7558.547	1200
16×16	6	7515.384	1536
	8	8480.487	2048
	12	9550.366	3072

Each frame is divided into a grid of $s \times s$ bins which is smaller than the size of frame. In addition to reducing the dimensionality size, this also accelerates the computing process. Different sizes of grids are investigated including 8×8 and 10×10 . The location of each feature is recorded, and the direction of the flow is categorized as one of the six motions $\{0, 60, 120, 180, 240, 300, 360\}$, eight motions $\{0, 45, 90, 135, 180, 225, 270, 315, 360\}$ or 12 motions $\{0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, 360\}$. The number of flows belongs to each direction is then counted to end up with an $8 \times 8 \times 6$, $8 \times 8 \times 8$, $8 \times 8 \times 10$, $8 \times 8 \times 12$, $10 \times 10 \times 6$, $10 \times 10 \times 8$, or $10 \times 10 \times 12$ bins descriptor for each frame. The average of the histograms in each grid for each video is calculated to come up with one feature vector. Table 5.4 shows the consumed time for visual feature extraction and the size of the extracted features (for the all videos in the dataset).

We performed several experiments using the previous settings, the best performance is obtained using $10 \times 10 \times 8$ descriptor (See Figure 5.5). Thus, in subsequent experiments each video is represented by a feature vector of size 800.

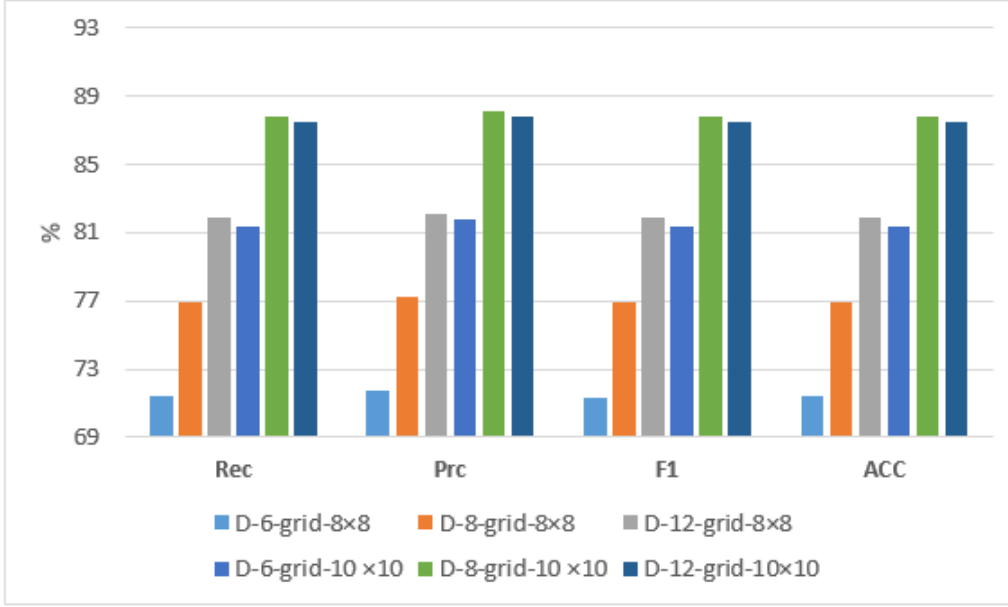


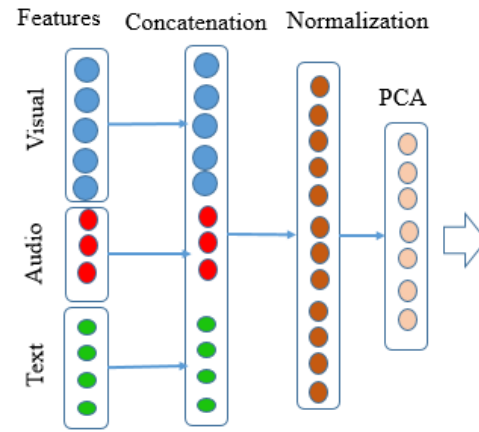
Figure 5.5: Results for different investigated parameters for visual features

5.4 Multimodal Sentiment Analysis

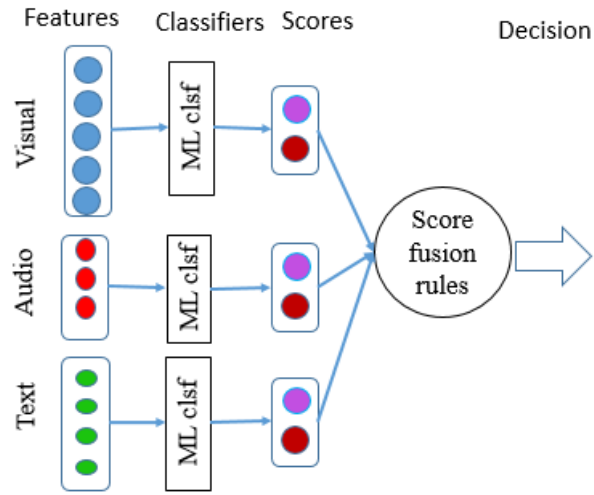
There are various directions to integrate multiple modalities depending on which modalities are chosen and at which level they are fused. Four main possibilities to combine Textual (T), Audio (A), and Visual (V) modalities, which are: A-T, T-V, A-V, and V-A-T. The fusion can be performed mainly at feature level, score level, decision level or hybrid.

5.4.1 Single-level Fusion

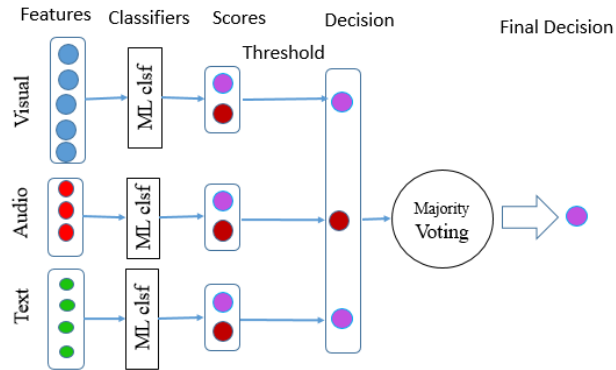
As a single-level fusion, three methods are explored and evaluated namely: feature-level (Figure 5.6 (a)), score-level (Figure 5.6(b)) and decision-level (Figure 5.6(c)). All of these fusion levels are described in Section 4.2.3. Table 5.5 describes the feature vectors.



(a)



(b)



(c)

Figure 5.6: Single-level fusion techniques: (a) feature, (b) score, and (c) decision

Table 5.5: The size of feature vectors

Feature vector	Size
Textual	300
Audio	68
Visual	800
Audio-textual	368
Audio-visual	868
Textual-Visual	1100
Audio-textual-visual	1168

5.4.2 Multi-level Hybrid Fusion

The hybrid fusion is conducted by utilizing and combining aforementioned fusion levels. It might include the advantages of other fusion levels when it is developed perfectly. Different possibilities can be designed using it such as fusing two modalities in feature level then fusing the resultant modality with other modalities in the score or decision level. A study of [95] applied the hybrid fusion level such that they combined the audio and visual modalities using the feature level fusion and the resultant model is combined with the textual modality using the decision level fusion.

In this study, two architectures for hybrid fusion are considered: (1) fusing feature level with score level and (2) fusing feature level with decision level. Combining these with the three modalities, we have the following systems:

- Audio and text modalities are fused at feature level (AT) and then AT is fused with audio and text modalities individually at score or decision level (AT-A-T).
- Text and visual modalities are fused at feature level (TV) and then TV is fused with text and visual modalities individually at score or decision level (TV-T-V).
- Audio and visual modalities are fused at feature level (AV) and then AV is fused with audio and visual modalities individually at score or decision level (AV-A-V).

- Audio and text modalities are fused at feature level (AT) and then AT is fused with audio, text and visual modalities individually at score or decision level (AT-A-T-V).
- Text and visual modalities are fused at feature level (TV) and then TV is fused with audio, text and visual modalities individually at score or decision level (TV-A-T-V).
- Audio and visual modalities are fused at feature level (AV) and then AV is fused with audio, text and visual modalities individually at score or decision level (AV-A-T-V).
- All modalities are combined at feature level (ATV) and then ATV is fused with audio, text and visual modalities individually at score or decision level (ATV-A-T-V).

5.5 Experiments and Results

5.5.1 Experimental Settings

Experiments are conducted on the developed Arabic multimodal dataset for sentiment analysis using 10-fold cross validation. For each instance in the dataset, an acoustic feature vector of 68 features is created for audio modality and a textual feature vector of 300 features is created for textual modality while a visual feature vector of 800 features is created for face expression modality. Two machine classifiers are applied to evaluate the proposed approaches, LibSVM with Linear Kernel and LR with L2

Table 5.6: Unimodal systems (Baseline).

Modality	SVM				LR			
	Rec	Prc	F_1	Acc	Rec	Prc	F_1	Acc
Audio	81.87	81.89	81.84	81.87	79.77	79.77	79.77	79.77
Text	82.06	82.06	82.05	82.06	83.40	83.40	83.38	83.40
Visual	88.17	88.22	88.17	88.17	85.69	85.75	85.69	85.69

norm regularization and Liblinear solver. Gensim package [231] is applied for textual features extraction, PayAudioAnalysis [232] package is utilized for acoustic features extraction while OpenCV [233] tool is utilized for visual features extraction. Scikit-learn package [160] is used for feature reduction, classification and evaluation. The imbalance class problem is also addressed through considering cost-sensitive classification. This is conducted through setting penalty parameter of the error term C_0 of class i as $C_1 = class_weight[i] * C_0$, where:

$$class_weight[i] = n_samples / (n_classes \times n_samples[i]) \quad (5.4)$$

5.5.2 Unimodal Results

The results of standalone audio, text and visual modalities are illustrated in Table 5.6. The highest results are in bold. Visual modality gives the highest results using SVM. Unlike audio and visual modalities, LR achieves higher results with text modality than SVM. Overall the average performance of SVM is higher than LR. The results archived using uni-modalities are considered the baseline for later comparisons of other systems.

5.5.3 Single-level Fusion Results

Several experiments are run to evaluate different early and late fusion techniques. The results of bi-modal and multimodal approaches using feature, score and decision level techniques are shown in Table 5.7. For feature level fusion, the highest results are obtained using multimodal approach of audio, text and visual (A-T-V) using LR. For score level fusion, the highest results are obtained using the multimodal approach (A-T-V) using product fusion rule with SVM. One case is considered for decision level to combine the three modalities. This is because such fusion technique requires odd number of modalities and applying decision fusion level with bi-modal approaches is meaningless using majority voting. Combining different modalities leads to improving the results of the unimodal approaches significantly for all cases. In general, single-level score fusion achieves the highest results with an improvement of around 5% than the highest results obtained using uni-modalities (visual modality with SVM).

5.5.4 Multi-level Hybrid Fusion Results

Different techniques are evaluated in order to fuse two and three modalities in multiple levels. In the first level, feature fusion technique is used to combine the considered modalities while in the second level either score or decision level fusion. The aim of multi-level hybrid fusion is to take the advantage of different fusion techniques in order to enhance the results. The results of multi-level fusion techniques are presented in Table 5.8. The highest results are obtained when fusing audio and visual modalities in the first level using feature fusion and the resultant modality are fused with audio,

Table 5.7: Single-level fusion of feature, score and decision techniques

Modality	Fusion	SVM				LR			
		Rec	Prc	F_1	Acc	Rec	Prc	F_1	Acc
Audio, Text	A-T	86.83	86.83	86.83	86.83	88.36	88.36	88.36	88.36
	sum(A-T)	88.36	88.57	88.32	88.36	89.31	89.35	89.3	89.31
	prod(A-T)	88.36	88.57	88.32	88.36	89.31	89.35	89.3	89.31
	max(A-T)	88.36	88.57	88.32	88.36	89.31	89.35	89.3	89.31
Text, Visual	T-V	89.31	89.35	89.3	89.31	87.6	87.61	87.6	87.60
	sum(T-V)	90.46	90.46	90.45	90.46	91.22	91.23	91.22	91.22
	prod(T-V)	90.46	90.46	90.45	90.46	91.22	91.23	91.22	91.22
	max(T-V)	90.46	90.46	90.45	90.46	91.22	91.23	91.22	91.22
Audio, Visual	A-V	92.56	92.57	92.56	92.56	90.46	90.48	90.45	90.46
	sum(A-V)	90.65	90.67	90.64	90.65	90.08	90.1	90.07	90.08
	prod(A-V)	90.65	90.67	90.64	90.65	90.08	90.1	90.07	90.08
	max(A-V)	90.65	90.67	90.64	90.65	90.08	90.1	90.07	90.08
Audio, Text, Visual	A-T-V	92.56	92.58	92.56	92.56	92.75	92.75	92.75	92.75
	sum(A-T-V)	93.13	93.16	93.12	93.13	92.75	92.76	92.74	92.75
	prod(A-T-V)	93.32	93.36	93.31	93.32	92.37	92.37	92.36	92.37
	max(A-T-V)	91.79	91.85	91.78	91.79	92.37	92.37	92.36	92.37
	mode(A-T-V)	92.94	92.96	92.93	92.94	90.84	90.84	90.84	90.84

text and visual modalities in the second level using score fusion technique (AV-A-T-V) using product rule with SVM classifier. In general multi-level hybrid fusion achieves the highest results comparing to unimodal approaches and single-level fusion approaches.

5.5.5 Enhancement of Visual Features

The idea is to combine local and global descriptors. Oriented FAST and Rotated BRIEF (ORB) [234] technique is applied to extract local descriptors from each frame in a video. ORB is a hybrid modified version of FAST key-point detector and BRIEF descriptor. First, FAST is applied to find key points, then Harris corner measure is used to find top N points among them. In order to produce multi-scale-features, pyramid is used. For computing orientation, ORB calculates the intensity weighted centroid of the patch with located corner at center. The orientation can be found as

Table 5.8: Multi-level hybrid fusion of feature, score, and decision fusion techniques

Modality	Fusion	SVM				LR			
		Rec	Prc	F_1	Acc	Rec	Prc	F_1	Acc
Audio, Text	sum(AT-A-T)	88.36	88.40	88.34	88.36	88.74	88.74	88.74	88.74
	prod(AT-A-T)	88.74	88.77	88.73	88.74	88.74	88.74	88.74	88.74
	max(AT-A-T)	88.17	88.22	88.15	88.17	89.31	89.32	89.31	89.31
	mode(AT,A,T)	87.79	87.79	87.78	87.79	88.17	88.17	88.17	88.17
Text, Visual	sum(TV-T-V)	91.03	91.03	91.03	91.03	90.84	90.85	90.84	90.84
	prod(TV-T-V)	91.03	91.05	91.02	91.03	90.84	90.85	90.84	90.84
	max(TV-T-V)	91.03	91.07	91.02	91.03	91.03	91.03	91.03	91.03
	mode(TV,T,V)	90.65	90.65	90.65	90.65	88.74	88.75	88.74	88.74
Audio, Visual	sum(AV-A-V)	91.98	92.02	91.99	91.98	90.46	90.55	90.44	90.46
	prod(AV-A-V)	92.37	92.38	92.37	92.37	90.27	90.34	90.25	90.27
	max(AV-A-V)	91.41	91.41	91.41	91.41	90.27	90.32	90.25	90.27
	mode(AV,A,V)	92.75	92.76	92.75	92.75	90.27	90.28	90.26	90.27
Audio, Text, Visual	sum(AT-A-T-V)	92.94	93.03	92.93	92.94	91.98	91.98	91.98	91.98
	prod(AT-A-T-V)	92.94	92.98	92.93	92.94	91.98	91.98	91.98	91.98
	max(AT-A-T-V)	91.79	91.85	91.78	91.79	92.56	92.56	92.56	92.56
	mode(AT,A,T,V)	90.27	90.62	90.27	90.27	89.89	90.29	89.88	89.89
Audio, Text, Visual	sum(TV-A-T-V)	92.18	92.18	92.17	92.18	91.98	92.00	91.99	91.98
	prod(TV-A-T-V)	92.37	92.38	92.36	92.37	92.18	92.19	92.18	92.18
	max(TV-A-T-V)	91.98	92.03	91.97	91.98	92.37	92.37	92.36	92.37
	mode(TV,A,T,V)	92.56	92.76	92.56	92.56	90.27	90.89	90.26	90.27
Audio, Text, Visual	sum(AV-A-T-V)	93.89	93.90	93.89	93.89	92.37	92.37	92.37	92.37
	prod(AV-A-T-V)	94.08	94.08	94.08	94.08	91.98	91.98	91.98	91.98
	max(AV-A-T-V)	92.94	92.95	92.94	92.94	92.75	92.75	92.75	92.75
	mode(AV,A,T,V)	93.32	93.57	93.32	93.32	91.79	92.09	91.80	91.79
Audio, Text, Visual	sum(ATV-A-T-V)	93.51	93.51	93.51	93.51	92.37	92.37	92.37	92.37
	prod(ATV-A-T-V)	93.13	93.14	93.13	93.13	92.56	92.56	92.56	92.56
	max(ATV-A-T-V)	92.75	92.76	92.74	92.75	92.18	92.18	92.17	92.18
	mode(ATV,A,T,V)	92.75	92.97	92.75	92.75	91.60	91.88	91.61	91.60

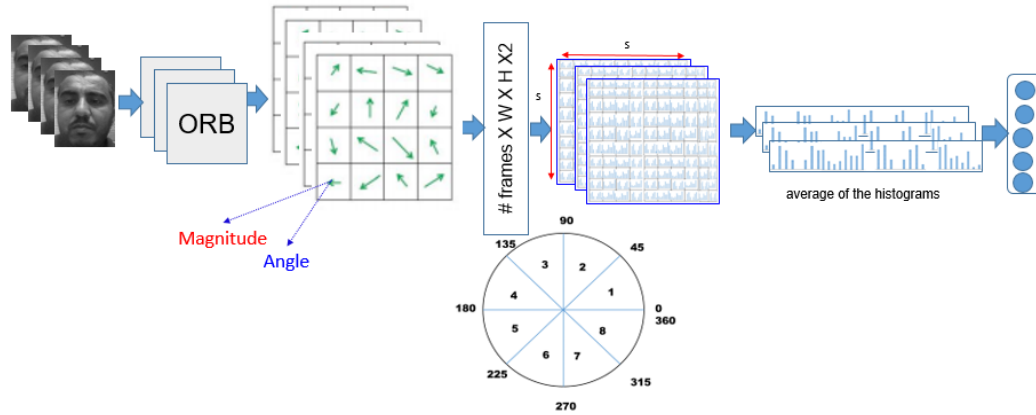


Figure 5.7: Combined ORB and dense optical flow features extraction process

Table 5.9: Multimodal sentiment analysis with the hybrid visual features

Fusion method	Modality	Rec	Prc	F_1	Acc
Single level	A-T-V	93.70	93.70	93.70	93.70
	sum(A-T-V)	93.70	93.80	93.69	93.70
	prod(A-T-V)	93.32	93.45	93.31	93.32
	max(A-T-V)	91.60	91.82	91.58	91.60
	mode(A,T,V)	93.13	93.24	93.12	93.13
Multi-level	sum(ATV-A-T-V)	95.23	95.23	95.23	95.23
	prod(ATV-A-T-V)	94.85	94.87	94.84	94.85
	max(ATV-A-T-V)	93.51	93.62	93.50	93.51
	mode(ATV,A,T,V)	94.08	94.12	94.09	94.08

the direction of the vector from this corner point to centroid. To improve the rotation invariance, moments are computed with x and y which should be in a circular region of radius r , where r is the size of the patch. The key points locations are drawn on each frame, then the histogram of dense optical flow (as global descriptors) is computed from the resultant image. Figure 5.7 depicts the hybrid of local and global visual feature extraction method. Using the same experimental settings and with SVM classifier, Table 5.9 presents the results of multimodal sentiment analysis with applying hybrid of local and global visual features for single and multi-level fusion. There are improvements in the results presented in Tables 5.7 and 5.8 related to visual features in case of multimodal sentiment analysis for both single level and multi-level fusion combining visual features with other modalities comparing to the results of single-level fusion reported in Table 5.7 and multi-level fusion reported in and Table 5.8. Nearly all results are improved remarkably except a case of $\max(\text{A-T-V})$ where there is negligible drop in the results. The highest performance are obtained with hybrid visual features in case of multi-level fusion of feature level and sum score level.

A	T	V	AT	TV	AV	ATV	max (A-T)	max (T-V)	max (A-V)	max (A-T-V)	mode (A-T-V)	prod (AV,A,T,V)	prod (A-T-V)	sum (A-T-V)
81.89	83.00	87.50	88.01	88.85	91.45	92.73	89.03	91.18	90.60	92.34	92.29	94.00	93.49	93.20
0.49	0.62	0.68	0.85	0.84	0.68	0.86	0.61	0.97	0.50	0.74	0.75	0.50	0.49	0.62
6.26E-05	1.00E-09	4.07E-09	1.44E-09	1.47E-10	6.83E-11	3.72E-11	3.88E-10	2.72E-11	1.31E-12	2.61E-12	6.90E-12	4.93E-12	1.09E-12	1.63E-12
	4.26E-08	6.02E-09	2.99E-08	2.27E-09	1.12E-09	1.12E-09	3.88E-10	7.88E-11	3.05E-11	3.54E-11	1.94E-11	5.15E-11	1.62E-11	5.39E-12
			2.46E-01	3.60E-03	1.12E-07	5.47E-07	1.36E-03	7.09E-07	9.01E-07	1.08E-07	5.14E-09	4.73E-09	1.06E-09	1.94E-09
				4.37E-02	7.24E-06	4.03E-07	3.75E-04	4.09E-05	1.50E-05	3.05E-07	1.99E-06	2.10E-08	5.30E-08	3.63E-08
					7.07E-05	5.26E-08	5.22E-01	8.55E-05	1.10E-04	5.05E-09	7.92E-06	2.81E-08	7.61E-09	1.32E-08
						8.51E-03	5.66E-05	6.01E-01	3.26E-02	4.94E-02	4.84E-02	1.93E-06	1.19E-04	6.02E-04
							9.72E-07	3.98E-03	6.59E-05	9.81E-02	3.55E-01	3.70E-04	1.31E-02	1.45E-01
								1.89E-04	1.63E-04	2.17E-07	5.42E-06	5.28E-08	1.67E-08	3.94E-08
									4.04E-02	9.50E-04	3.04E-03	4.86E-05	1.56E-05	2.67E-05
										2.26E-05	6.81E-05	5.50E-08	3.75E-07	1.17E-06
											8.91E-01	1.37E-04	2.57E-04	3.10E-03
												1.18E-03	1.93E-03	2.82E-03
													5.27E-02	2.01E-02
														4.51E-02
														prod (AV,A,T,V)
														prod (A-T-V)

Figure 5.8: p-values of pairwise t-test of sentiment analysis system

5.5.6 Analysis and Discussion

Statistical tests are conducted to explore whether the performance differences are significant or just achieved by chance. We re-run the experiments 10 times for each 10-fold classification model and used multiple pairwise two-sided t-test with 95% confidence interval (See Appendix B). We formulated the null hypothesis as follows:

$H_{0(X,Y)}$: there is no significant difference between system X and system Y ,

where X and Y can be unimodal, bi-modal or multi-modal system. The p-values for pairwise t-test on F_1 scores are shown in Figure 5.8 where the cases in which no significant differences (i.e., $p\text{-value} > 0.05$) are represented in bold.

The considered modalities to answer the first research question (RQ1) are Audio, Textual and Visual individually. It is observed that visual modality significantly performs better than other single modalities whereas textual modality significantly achieves higher results than audio modality.

To analyze the effects of combining different modalities (RQ2), we considered audio, textual and visual modalities with different possibilities to combine them. It is observed that combining different modalities improves the results significantly than uni-modalities. In addition, tri-modalities significantly perform better than the bi-modalities.

To answer the research question (RQ3), we tested feature level fusion, score level fusion using PROD rule and decision level fusion. It is observed that there is no significant difference between feature, score and decision fusion methods using single level fusion. To evaluate the most efficient score fusion method, we considered MAX,

Table 5.10: Benchmarking our results

Approach	Audio	Text	Visual	Dataset	Analysis level	Fusion	Best performance
[126]	✓		✓	21 videos	video	feature level	Acc: 85
[123]		✓	✓	40 videos	utterance	feature level	Acc: 76.09
Ours	✓	✓	✓	63 videos	utterance	feature, score, decision, hybrid	Acc: 95.23

PROD and SUM score fusion rules. First, it should be emphasized that in case of combining two modalities with binary class, MAX, PROD and SUM rules achieve the same performance. However, in case of multi-modality, the PROD rule significantly performs better than the MAX and SUM rules whereas the SUM rule significantly outperforms the MAX rule.

To analyze the effects of hybrid multi-level fusion comparing with the single level fusion (RQ4), we considered feature level fusion, score level fusion and decision level fusion as single level with AV-A-T-V as hybrid multi-level fusion. Hybrid fusion level significantly outperforms feature fusion level and decision fusion level.

Table 5.10 benchmarks our work with the two most related work in the literature on multimodal Arabic sentiment analysis [123, 126]. Our work compares favorably with the other works in several aspects. It is the first study to address multimodal Arabic sentiment analysis considering three modalities. It develops larger and comprehensive dataset for sentiment analysis and demographic detection. It also evaluates different features and several fusion methods. Furthermore, it achieves higher results.

5.6 Summary

The experimental results show that:

- Multimodal sentiment analysis approach is more accurate than stand-alone ones

and improves the unimodal approaches significantly in all cases.

- Although score level fusion achieves higher results than feature and decision levels, there is no significant difference among them in case of single-fusion level and multimodal system.
- The PROD rule score fusion significantly outperforms the MAX and SUM rules.
- The proposed multi-level fusion method of feature level with score level reports the highest results.
- The proposed hybrid visual features improves the results in case of multimodal sentiment analysis.

CHAPTER 6

EFFECTS OF DEMOGRAPHICS ON SENTIMENT ANALYSIS

Demographic analysis refers to studying the composition characteristics of a group of people. These characteristics may include gender, age, race (ethnicity), dialect (accent), education level, disability, household income and nationality. In this study, we only considered three characteristics (gender, age and dialect) as examples but the model can be generalized to cover more characteristics. As mentioned in Section 1.1, combining demographics with sentiment analysis can lead to better understanding of people's opinions and personalization of services. For example, large companies can enhance and improve their products and/or services to a group of customers based on their demographics. Educational and training systems can be more adaptive and interactive where the content is tailored to the learners' emotion, age-groups, genders, or dialects. The research and the resources in this direction are scarce in the Arabic language.

In this chapter, we present and discuss the results of several experiments using our annotated dataset and feature extraction, as described in Sections 5.2 and 5.3, for unimodal, bimodal, and trimodal demographic and sentiment analysis from opinion videos. In Section 6.1, we study the demographic characteristics of our dataset and present the results for machine learning classification of gender, age-group and dialect. Section 6.2 presents the results for multi-class classification models that simultaneously recognize demographics as well as sentiment. Section 6.3 presents the results for multi-label multi-class systems for demographics and sentiment classification. In Section 6.4, we study the impact of demographics, taken as known inputs, on sentiment classification. In our experiments, we used SVM classifier and 10-fold cross-validation, and employed feature level fusion to integrate various modalities. Finally, in Section 6.5 we explore an enhanced multi-class multimodal demographic and sentiment detection system taken into consideration new features computed using Histogram of Oriented Gradient (HOG). For multi-class classification, one versus rest/all strategy is applied. We also conducted statistical tests to compare different models.

6.1 Demographic Detection

Several experiments are first conducted for each demographic characteristic and the average results are shown in Table 6.1. Besides several performance measures (Rec, Prc, F_1 , GM, Acc, MCC), it also shows the training (fitting) time (T_f), the testing (scoring) time (T_s) and the total time ($T_t = T_f + T_s$). It can be observed that for

Table 6.1: Results for demographic recognition systems

Demographic	Modality	Rec	Prc	F_1	GM	Acc	MCC	T_f	T_s	T_t
Gender	Audio	95.61	95.61	95.61	95.37	95.61	90.93	2.13	0.22	2.35
	Text	75.00	74.79	74.75	73.39	75.00	47.78	9.05	0.69	9.74
	Visual	90.65	90.63	90.64	90.25	90.65	80.67	20.08	1.32	21.40
	A-T	92.37	92.36	92.36	91.98	92.37	84.21	24.29	1.15	25.44
	T-V	91.22	91.21	91.21	90.80	91.22	81.84	48.00	4.38	52.38
	A-V	95.80	95.80	95.80	95.53	95.80	91.32	38.05	3.20	41.25
	A-T-V	95.42	95.44	95.43	95.41	95.42	90.59	43.61	3.96	47.57
Age	Audio	69.47	69.50	69.31	78.64	69.47	58.78	7.26	0.50	7.76
	Text	53.82	53.63	53.70	67.05	53.82	37.82	24.22	1.54	25.76
	Visual	74.81	75.08	74.84	82.50	74.81	66.10	69.28	3.82	73.10
	A-T	73.47	73.48	73.44	81.56	73.47	64.20	32.18	3.28	35.46
	T-V	75.76	75.78	75.77	83.25	75.76	67.32	55.30	4.78	60.08
	A-V	84.35	84.48	84.39	89.27	84.35	78.93	50.35	4.11	54.46
	A-T-V	86.64	86.66	86.65	90.87	86.64	81.99	61.53	5.18	66.71
Dialect	Audio	71.56	71.48	71.34	76.42	71.56	55.35	5.71	0.42	6.13
	Text	72.33	71.29	70.94	77.47	72.33	55.85	16.71	1.04	17.75
	Visual	83.97	83.97	83.79	86.71	83.97	74.98	62.41	3.23	65.64
	A-T	85.69	85.69	85.52	88.56	85.69	77.59	27.84	2.89	30.73
	T-V	87.98	88.17	87.88	90.15	87.98	81.28	50.47	5.47	55.94
	A-V	88.36	88.44	88.28	90.21	88.36	81.87	50.08	5.04	55.12
	A-T-V	90.65	90.79	90.56	92.19	90.65	85.42	58.95	5.93	64.88

multi-class problems (age-group and dialect), the multimodal system A-T-V has the highest performance, e.g., the F_1 score reaches about 86.65% for age-group and 90.56% for dialect. However, compared to single modalities there has been an increase in the time complexity during training and testing but it remains around one minute for training and about five seconds for testing. The corresponding confusion matrices are shown in Figure 6.1.

Statistical tests are conducted to explore whether the performance differences are significant or just achieved by chance. We re-run the experiments 10 times for each 10-fold classification model and used multiple pairwise two-sided t-test with 95% confidence interval. We formulated the null hypothesis as follows:

$H_{0(X,Y)}$: there is no significant difference between system X and system Y ,

where X and Y can be any of the seven systems (A, T, V, A-T, T-V, A-V and A-T-V).

		Predicted		
		Fernal	Male	Total
Actual	Male	294	14	308
	Female	10	206	216
	Total	304	220	524

(a)

		Predicted				
		AGA	AGB	AGC	AGD	Total
Actual	AGA	114	4	7	3	128
	AGB	4	141	9	5	159
	AGC	5	13	118	6	142
	AGD	3	2	9	81	95
	Total	126	160	143	95	524

(b)

		Predicted				
Actual		Egyptian	Gulf	LEV	Maghribi	Total
	Egyptian	239	9	1	0	249
	Gulf	16	148	3	0	167
	LEV	7	10	62	0	79
	Maghribi	2	1	0	26	29
	Total	264	168	66	26	524

(c)

Figure 6.1: Confusion matrix for demographic detection systems: (a) Gender, (b) Age-group, and (c) Dialect

Demographic	A	T	V	AT	TV	AV	ATV	
Gender	94.68	74.62	90.45	92.51	91.53	96.45	96.34	AVG
	2.75	6.22	3.67	3.17	3.41	2.37	2.40	STD
		1.85E-51	7.70E-14	1.79E-08	2.37E-11	4.20E-06	7.82E-06	A
			2.14E-38	5.48E-51	3.19E-44	4.67E-54	1.48E-58	T
				1.15E-04	2.79E-03	1.52E-30	9.43E-30	V
					3.52E-02	5.97E-17	1.76E-19	AT
						1.99E-22	1.35E-25	TV
							6.96E-01	AV
Age-group	A	T	V	AT	TV	AV	ATV	
	69.22	54.41	76.08	71.45	75.94	84.98	86.24	AVG
	5.25	7.21	6.12	6.02	5.49	4.41	4.64	STD
		1.46E-30	1.82E-14	1.72E-03	2.83E-15	2.66E-45	3.09E-47	A
			3.66E-42	5.45E-46	1.61E-48	7.08E-59	1.36E-64	T
				9.05E-08	7.60E-01	1.60E-33	9.61E-34	V
					4.54E-09	5.93E-36	1.25E-43	AT
						5.82E-34	5.84E-38	TV
							8.98E-04	AV
Dialect	A	T	V	AT	TV	AV	ATV	
	71.71	72.06	82.94	84.39	84.59	88.41	90.63	AVG
	5.90	5.20	5.40	4.21	4.92	4.24	3.63	STD
		6.64E-01	1.86E-27	1.96E-40	5.03E-32	1.16E-45	6.63E-50	A
			1.88E-28	5.07E-42	1.13E-33	1.33E-46	3.82E-57	T
				3.47E-02	1.08E-03	1.32E-20	9.44E-26	V
					7.56E-01	1.43E-10	1.47E-22	AT
						4.48E-12	8.04E-22	TV
							9.04E-06	AV

Figure 6.2: p-values of pairwise t-test of demographics

The p-values for pairwise t-test on F_1 scores are shown in Figure 6.2 where the cases in which no significant differences (i.e., p -value > 0.05) are represented in bold.

6.2 Multi-class Demographic with SA

In the following, each of the considered demographics is investigated in combination with sentiment. This includes the tasks of (1) detecting gender with sentiment (G-S) (i.e., 4 binary variable as shown in Figure 6.3), (2) detecting age-group with sentiment (A-S) (i.e., 8 binary variables), (3) detecting dialects with sentiment (D-S) (i.e., 8 binary variables), (4) detecting gender and age-group with sentiment (G-A-S) (i.e.,

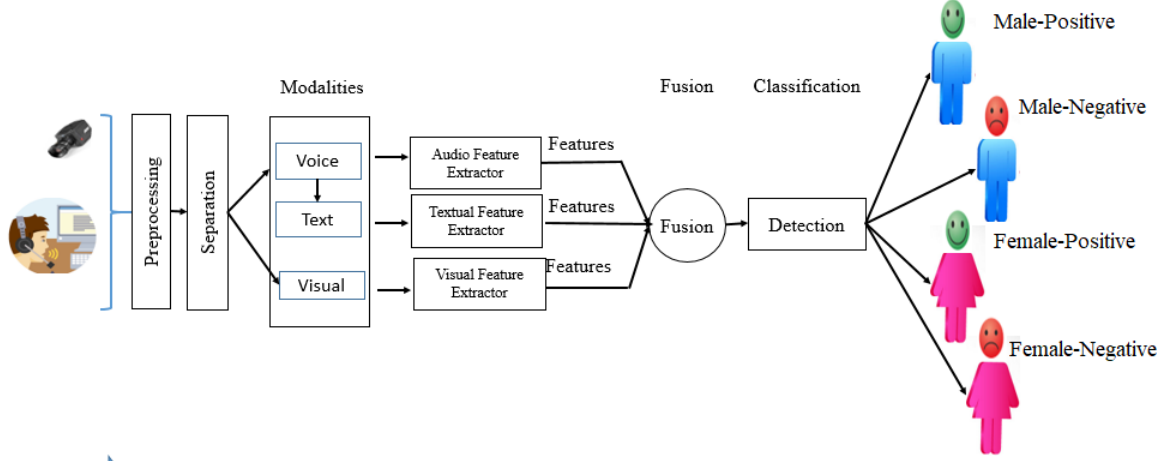


Figure 6.3: Layout of the process for multi-class gender with sentiment detection

Table 6.2: Results for multi-class demographic with sentiments recognition systems

Demographic	Modality	Rec	Prc	F_1	GM	Acc	MCC	T_f	T_s	T_t
G-S	Audio	77.48	77.26	77.08	84.23	77.48	69.31	4.80	0.38	5.18
	Text	66.03	65.19	65.22	75.98	66.03	53.51	18.26	1.16	19.42
	Visual	86.07	86.05	86.01	90.28	86.07	81.05	55.35	3.27	58.62
	A-T	81.87	81.81	81.82	87.62	81.87	75.34	26.96	2.50	29.46
	T-V	85.69	85.70	85.68	90.13	85.69	80.55	45.39	4.13	49.52
	A-V	90.84	90.87	90.82	93.66	90.84	87.56	43.00	3.22	46.22
	A-T-V	91.03	91.08	90.97	93.80	91.03	87.82	51.48	4.06	55.54
A-S	Audio	71.37	71.35	71.03	81.93	71.37	65.73	8.19	0.58	8.77
	Text	55.34	54.68	53.93	70.75	55.34	46.08	27.05	1.94	28.99
	Visual	78.44	78.85	78.13	86.54	78.44	74.20	79.57	4.49	84.06
	A-T	71.76	71.94	71.61	82.13	71.76	66.13	38.11	3.02	41.13
	T-V	79.01	79.14	78.96	86.97	79.01	74.90	51.76	4.96	56.72
	A-V	86.83	86.98	86.83	91.87	86.83	84.29	59.11	5.27	64.38
	A-T-V	88.36	88.56	88.37	92.89	88.36	86.12	67.46	6.86	74.32
D-S	Audio	71.76	71.85	71.56	81.62	71.76	65.22	6.66	0.54	7.20
	Text	64.89	65.64	64.30	76.70	64.89	56.49	22.67	1.53	24.20
	Visual	81.87	81.88	81.66	88.33	81.87	77.74	69.53	3.71	73.24
	A-T	80.53	80.80	80.56	87.50	80.53	76.11	31.98	2.10	34.08
	T-V	85.50	85.63	85.42	90.64	85.50	82.20	49.35	4.21	53.56
	A-V	90.27	90.22	90.14	93.86	90.27	88.12	51.00	4.43	55.43
	A-T-V	93.13	93.18	93.11	95.66	93.13	91.60	64.11	5.78	69.89
G-A-S	Audio	75.57	76.13	75.10	85.71	75.57	73.08	11.00	0.89	11.89
	Text	55.92	55.06	54.71	72.90	55.92	51.13	37.85	2.56	40.41
	Visual	84.54	84.82	84.07	91.21	84.54	82.99	131.83	6.58	138.41
	A-T	79.01	79.37	78.97	87.91	79.01	76.88	45.63	4.50	50.13
	T-V	87.98	88.01	87.81	93.21	87.98	86.78	84.54	6.80	91.34
	A-V	92.56	92.67	92.52	95.81	92.56	91.81	97.41	7.03	104.44
	A-T-V	92.37	92.45	92.33	95.69	92.37	91.60	118.40	9.29	127.69

16 binary variables). They are evaluated using the developed dataset in case of uni-modality, bi-modality and multi-modality with the same experimental settings. The average results are shown in Table 6.2 and the corresponding confusion matrices are shown in Figures 6.4 and 6.5. The p-values for pairwise t-test on F_1 scores of the demographic systems are shown in Figure 6.2 where the cases in which no significant differences are represented in bold. It can be observed that for A-S and D-S, the multimodal system A-T-V has significantly higher performance than other systems. For G-S and G-A-S although A-V and A-T-V significantly perform better than other systems, there is not much difference between them. However, this comes at a slight increase in the consumed time.

6.3 Multi-label Demographic and SA

Here, we show the results for a monolithic multi-label model trained once and able to detect sentiment (positive or negative), gender (male or female), and four age-groups (A, B, C or D) and their different combinations. The performance measures of the system are analyzed for each individual label (Sentiment, Gender and Age), pairs of labels (Sentiment-Gender, Sentiment-Age, Gender-Age), and for the three labels together (Sentiment-Gender-Age). Figures 6.7 shows the confusion matrices for each target variable: sentiment, gender, and age-group. Clearly the system can distinguish with high accuracy between various classes in each case. The problem of multi-label classification is more challenging than binary classification and that is why the system has slightly decreased accuracy in the case of age-groups due to the existence of four

		Predicted				
		Male-Neg	Male-Pos	Female-Neg	Female-Pos	Total
Actual	Male-Neg	157	5	3	3	168
	Male-Pos	9	127	0	4	140
	Female-Neg	8	1	66	7	82
	Female-Pos	1	2	4	127	134
	Total	175	135	73	141	524

(a)

Actual	Predicted								Total
	AGA-Neg	AGA-Pos	AGB-Neg	AGB-Pos	AGC-Neg	AGC-Pos	AGD-Neg	AGD-Pos	
	AGA-Neg	20	3	0	0	2	0	0	25
	AGA-Pos	0	93	0	3	2	1	4	103
	AGB-Neg	0	1	42	1	1	0	1	46
	AGB-Pos	0	1	2	100	5	2	2	113
	AGC-Neg	0	1	1	3	81	1	8	95
	AGC-Pos	0	1	0	1	1	44	0	47
	AGD-Neg	0	3	1	1	5	1	72	84
	AGD-Pos	0	0	0	0	0	0	11	11
	Total	20	103	46	109	97	49	87	524

(b)

Actual	Predicted								Total
	Egypt-Neg	Egypt-Pos	Gulf-Neg	Gulf-Pos	Lev-Neg	Lev-Pos	Magh-Neg	Magh-Pos	
	Egypt-Neg	138	3	0	1	0	1	0	143
	Egypt-Pos	9	92	2	2	0	1	0	106
	Gulf-Neg	2	0	57	2	0	0	0	61
	Gulf-Pos	0	3	2	98	0	3	0	106
	Lev-Neg	0	0	0	0	32	1	0	33
	Lev-Pos	1	1	0	2	0	42	0	46
	Magh-Neg	0	0	0	0	0	0	13	13
	Magh-Pos	0	0	0	0	0	0	16	16
	Total	150	99	61	105	32	48	13	524

(c)

Figure 6.4: Confusion matrix of multi-class demographic with SA: (a) Gender-sentiment, (b) Age-group-sentiment, and (c) Dialect-sentiment

		Predicted											Total
	Actual	Male-AGA-Neg	Male-AGA-Pos	Male-ACB-Neg	Male-ACB-Pos	Male-ACC-Neg	Male-ACC-Pos	Male-ACD-Neg	Male-ACD-Pos	Male-AGA-Pos	Male-ACB-Neg	Male-ACB-Pos	
		21	1	0	0	0	0	1	0	2	0	0	25
Male-AGA-Pos		0	36	0	1	1	1	1	0	0	0	2	41
Male-ACB-Neg		0	0	22	0	0	0	0	0	0	0	0	22
Male-ACB-Pos		1	0	0	73	0	0	1	0	0	1	0	77
Male-ACC-Neg		0	0	0	3	42	0	6	0	0	0	1	52
Male-ACC-Pos		0	0	0	1	0	15	0	0	0	0	0	16
Male-ACD-Neg		0	1	0	1	3	0	62	0	0	0	1	69
Male-ACD-Pos		0	0	0	0	0	0	0	6	0	0	0	6
Female-AGA-Pos		0	0	0	0	0	0	0	0	61	0	0	62
Female-ACB-Neg		0	0	0	0	0	0	0	0	0	0	0	24
Female-ACB-Pos		0	0	0	0	0	0	0	0	23	0	1	24
Female-ACC-Pos		0	0	0	0	1	0	0	0	0	0	2	36
Female-ACD-Pos		0	0	0	0	0	0	1	0	0	0	0	43
Female-AGC-Neg		0	0	0	0	0	0	0	0	0	42	0	43
Female-AGC-Pos		0	0	0	0	0	0	0	0	0	0	30	31
Female-AGD-Neg		0	0	0	0	0	0	0	0	1	0	0	15
Female-AGD-Pos		0	0	0	0	0	0	0	0	0	0	0	5
Total		22	38	22	79	47	15	72	6	64	24	36	524

Figure 6.5: Confusion matrix of multi-class gender, age-group and sentiment

G-S	A	T	V	AT	TV	AV	ATV	
	77.08488	65.92373	86.11522	81.14971	87.23161	91.63427	91.78231	AVG
	5.546011	6.34466	5.229924	5.09009	5.030923	4.024874	3.975688	STD
		5.4E-24	9.4E-22	1.6E-10	2.9E-25	4.5E-43	1.0E-38	A
			1.5E-43	6.1E-45	3.7E-48	3.6E-56	2.6E-62	T
				5.2E-10	5.4E-03	5.2E-20	2.0E-19	V
					1.8E-13	3.7E-29	1.9E-32	AT
						9.2E-16	3.9E-16	TV
							7.1E-01	AV
A-S	A	T	V	AT	TV	AV	ATV	
	71.95	53.00	75.43	72.58	77.86	85.19	86.81	AVG
	6.74	6.78	5.07	5.89	5.47	4.58	4.58	STD
		3.0E-37	1.3E-04	4.1E-01	1.0E-09	2.4E-29	2.0E-34	A
			1.1E-43	3.4E-52	1.3E-47	3.0E-60	1.0E-68	T
				7.4E-04	6.7E-07	7.2E-37	3.5E-38	V
					2.3E-09	1.3E-31	3.1E-41	AT
						1.5E-24	1.6E-28	TV
							9.2E-04	AV
D-S	A	T	V	AT	TV	AV	ATV	
	72.16	63.67	81.56	79.41	84.27	89.39	93.09	AVG
	6.02	6.35	4.55	6.24	5.38	4.34	3.84	STD
		8.6E-20	4.5E-23	3.9E-15	2.7E-27	1.5E-44	1.2E-47	A
			5.3E-40	1.2E-46	4.5E-45	2.5E-56	5.0E-64	T
				4.5E-03	4.6E-08	2.7E-35	5.1E-43	V
					3.2E-08	5.5E-25	7.6E-36	AT
						1.4E-19	9.5E-37	TV
							1.8E-14	AV
G-A-S	A	T	V	AT	TV	AV	ATV	
	74.46	52.35	84.45	76.86	86.61	92.40	92.85	AVG
	5.97	6.94	4.86	5.23	4.65	3.83	3.39	STD
		1.8E-44	1.0E-25	1.1E-03	8.2E-31	7.3E-48	1.9E-48	A
			2.3E-63	6.4E-64	1.3E-66	6.3E-75	4.3E-77	T
				2.1E-21	1.6E-07	4.5E-31	1.8E-29	V
					2.9E-28	2.3E-47	4.2E-50	AT
						1.3E-21	1.6E-22	TV
							2.2E-01	AV

Figure 6.6: p-values of pairwise t-test of multi-class demographics with sentiments

classes.

Figures 6.8 shows the confusion matrices for each pair of labels: Sentiment-Gender, Sentiment-Age, and Gender-Age, respectively. The notation X_Y denotes the first label X and second label Y. For example, N_A means negative (N) sentiment and age-group A. Again we can observe here that the system has correctly classified most of the instances but sometimes few instances are confused to be of other types.

Finally, we considered all the three labels together and evaluated the potential of the system to correctly recognize the tri-labels: Sentiment-Gender-Age. The total number of combinations of the three labels are 16 (which is $2 \times 2 \times 4$). The results are shown in Figure 6.9. The majority of instances are classified on the diagonal of the matrix, which indicates that they are correctly classified. Some cases have low number of instances and the system gives them less priority in favor of major cases. For instance, case N_F_A has zero instance and P_F_D has five instances. By contrast, the case N_M_D has 69 instances of which the system correctly recognized 60.

Additionally, we computed the per-class performance for the single label cases, pair of labels cases and tri-label cases. The results are shown in Figure 6.10. The accuracy and average per class performance is also depicted in Figure 6.11. This demonstrates that the developed system has remarkable performance for single labels. However, it starts to decrease when we consider combination of labels. This makes sense since there are more cases that can be confused together. The proposed multimodal system reports the highest accuracy for gender detection followed by sentiment detection and sentiment-gender detection.

A) Sentiment

		Actual		
		N	P	Total
Predicted	N	232	17	249
	P	18	257	275
	Total	250	274	524

B) Gender

		Actual		
		F	M	Total
Predicted	F	206	9	215
	M	10	299	309
	Total	216	308	524

C) Age Groups

		Actual				
		A	B	C	D	Total
Predicted	A	111	4	6	3	124
	B	6	141	10	2	159
	C	9	9	118	8	144
	D	2	5	8	82	97
	Total	128	159	142	95	524

Figure 6.7: Confusion matrix for each label in multi-label system: (a) Sentiment only, (b) Gender only, and (c) Age-group only

6.4 Effects of Demographics

The objective of this part is to explore and evaluate the effectiveness of utilizing demographics as inputs when detecting sentiment with different modalities: audio, textual, visual and their combinations. Four characteristics are considered: gender, age-group, dialect and nationality. The experiments are conducted using the previous experimental settings, features, classifiers in case of feature level fusion. The results are presented in Table 6.3 which is divided into four parts:

- The first part presents the results of demographic modality. Around 75% is reported for all metrics excepts MCC which achieves 0.4924 (MCC values range from -1 to 1).
- The second part presents the results of using demographics with each of the

A) Sentiment_Gender

		Actual				
		N_F	N_M	P_F	P_M	Total
Predicted	N_F	70	2	6		78
	N_M	5	155		11	171
	P_F	7	1	123	6	137
	P_M		10	5	123	138
	Total	82	168	134	140	524

B) Sentiment_Age

		Actual								
		N_A	N_B	N_C	N_D	P_A	P_B	P_C	P_D	Total
Predicted	N_A	19	1	2	2	3				27
	N_B		41	7	2	1	4			55
	N_C	1	1	72	7	1	2	2		86
	N_D		1	7	69		1		3	81
	P_A	5			1	84	3	4		97
	P_B		2	0		5	94	3		104
	P_C			7		7	6	37	1	58
	P_D				3	2	3	1	7	16
	Total	25	46	95	84	103	113	47	11	524

C) Gender_Age

		Actual								
		F_A	F_B	F_C	F_D	M_A	M_B	M_C	M_D	Total
Predicted	F_A	50	1	3	2					56
	F_B	4	53	5			3			65
	F_C	6	3	60	3		1	3	1	77
	F_D		2	3	11				1	17
	M_A	2				59	3	3	1	68
	M_B		1			2	84	5	2	94
	M_C			2	1	3	5	53	3	67
	M_D			1	3	2	3	4	67	80
	Total	62	60	74	20	66	99	68	75	524

Figure 6.8: Confusion matrix for pair of labels in multi-label system: (a) Sentiment-Gender, (b) Sentiment-Age, and (c) Gender-Age

modalities (audio, textual and visual). Comparing with the results presented in Chapter 5, there are improvements for the audio modality by more than 4% in terms of all evaluation measures. This is also true regarding visual modality

		Actual																Total
		N_F_A	N_F_B	N_F_C	N_F_D	N_M_A	N_M_B	N_M_C	N_M_D	P_F_A	P_F_B	P_F_C	P_F_D	P_M_A	P_M_B	P_M_C	P_M_D	
Predicted	N_F_A		1	1	1					1								4
	N_F_B		19	2						1	1							23
	N_F_C			32	2			1	1		1	1						38
	N_F_D		1	3	8								1					13
	N_M_A					19		1	1					2				23
	N_M_B		1				21	5	2						3			32
	N_M_C			1	1	1	1	38	3					1	1	1		48
	N_M_D			1	1			3	60						1		2	68
	P_F_A				1					49		2						52
	P_F_B		2							3	31	3			3			42
	P_F_C			3				1		6	2	24	1		1	1		39
	P_F_D				1						1		1				1	4
	P_M_A					5				2				33	3	2		45
	P_M_B													2	60			62
	P_M_C							3				1		1	3	11		19
	P_M_D								2				2	2	2	1	3	12
	Total	0	24	43	15	25	22	52	69	62	36	31	5	41	77	16	6	524

Figure 6.9: Confusion matrix for tri-labels in multi-label system: Sentiment-Gender-Age

Category	Rec	Prc	F1
N_F	0.8537	0.8974	0.8750
N_M	0.9226	0.9064	0.9145
P_F	0.9179	0.8978	0.9077
P_M	0.8786	0.8913	0.8849
N_A	0.7600	0.7037	0.7308
N_B	0.8913	0.7455	0.8119
N_C	0.7579	0.8372	0.7956
N_D	0.8214	0.8519	0.8364
P_A	0.8155	0.8660	0.8400
P_B	0.8319	0.9038	0.8664
P_C	0.7872	0.6379	0.7048
P_D	0.6364	0.4375	0.5185
F_A	0.8065	0.8929	0.8475
F_B	0.8833	0.8154	0.8480
F_C	0.8108	0.7792	0.7947
F_D	0.5500	0.6471	0.5946
M_A	0.8939	0.8676	0.8806
M_B	0.8485	0.8936	0.8705
M_C	0.7794	0.7910	0.7852
M_D	0.8933	0.8375	0.8645

Category	Rec	Prc	F1
N_F_A	1.0000	0.0000	0.0000
N_F_B	0.7917	0.8261	0.8085
N_F_C	0.7442	0.8421	0.7901
N_F_D	0.5333	0.6154	0.5714
N_M_A	0.7600	0.8261	0.7917
N_M_B	0.9545	0.6563	0.7778
N_M_C	0.7308	0.7917	0.7600
N_M_D	0.8696	0.8824	0.8759
P_F_A	0.7903	0.9423	0.8596
P_F_B	0.8611	0.7381	0.7949
P_F_C	0.7742	0.6154	0.6857
P_F_D	0.2000	0.2500	0.2222
P_M_A	0.8049	0.7333	0.7674
P_M_B	0.7792	0.9677	0.8633
P_M_C	0.6875	0.5789	0.6286
P_M_D	0.5000	0.2500	0.3333

Category	Rec	Prc	F1
N	0.9280	0.9317	0.9299
P	0.9380	0.9345	0.9362
F	0.9537	0.9581	0.9559
M	0.9708	0.9676	0.9692
A	0.8672	0.8952	0.8810
B	0.8868	0.8868	0.8868
C	0.8310	0.8194	0.8252
D	0.8632	0.8454	0.8542

(a) Single label

(b) Pair of labels

(c) Tri-labels

Figure 6.10: Per-class performance in terms of precision, recall and F_1 measure w.r.t.: (a) Single label, (b) Pair of labels, and (c) Tri-labels

where it is improved by more than 2% in terms of all measures. However, textual modality performance is slightly dropped with less than 1% in terms of all evaluation measures.

To perform statistical test to explore the significance of combining demographic

Category	Rec	Prc	F1	Acc
Sentiment	0.9330	0.9331	0.9331	0.9332
Gender	0.9622	0.9629	0.9626	0.9637
Age	0.8620	0.8617	0.8618	0.8626
Sentiment_Gender	0.8932	0.8982	0.8955	0.8989
Sentiment_Age	0.7877	0.7479	0.7630	0.8073
Gender_Age	0.8082	0.8155	0.8107	0.8340
Sentiment_Gender_Age	0.7363	0.6572	0.6582	0.7805

Figure 6.11: Accuracy and average precision, recall and F_1 measure for Sentiment, Gender, Age, Sentiment-Gender, Sentiment-Age, Gender-Age, and Sentiment-Gender-Age

characteristics as features with different modalities. We re-run the 10-fold cross-validation 10 times and used the pairwise t-test with 95% confidence interval.

Combining text modality with demographic features leads to improve the results significantly with p-value of 0.0098. This is also true regarding audio modality and visual modality where combining each of them individually with demographic features leads to improve the results significantly with p-value of less than 0.00001. There is significant improvement when combining the single modalities with demographic features to detect the sentiment of the speakers.

- The third part presents the results of combining demographics with bi-modalities. Comparing with results presented in Chapter 5, there are improvements for the audio-visual modality by more than 1.5% in terms of all evaluation measures. This is also true regarding textual-visual modality where all measures are improved by around 3%. However, audio-textual modality has not affected. Combining audio-visual modality with demographic features leads to significantly improving the results (p-value of 0.0011). This is true as well regarding

Table 6.3: Results demographic as a new modality with its effects on other modalities

Modality	Rec	Prc	F_1	GM	Acc	MCC
Demo	74.62	74.69	74.63	74.64	74.62	49.24
A-D	85.88	85.89	85.86	85.78	85.88	71.69
T-D	81.49	81.48	81.48	81.42	81.49	62.88
V-D	90.46	90.46	90.45	90.40	90.46	80.87
A-T-D	86.83	86.83	86.83	86.81	86.83	73.61
A-V-D	92.56	92.58	92.56	92.59	92.56	85.11
T-V-D	92.18	92.20	92.18	92.20	92.18	84.35
A-T-V-D	94.66	94.66	94.66	94.65	94.66	89.29

textual-visual modality when combining with demographic features. However, combining demographic features with audio-textual modality drops the results. Thus, the null hypothesis is rejected in case of audio-visual and textual-visual and accepted in case of audio-textual modality.

- The fourth part presents the results of employing demographics with multi-modalities. There are improvements for the audio-textual-visual by around 1% . Combining demographic features with audio-textual-visual improves the results significantly with p-value of less than 0.00001.

6.5 Enhanced Multi-class Demographic with SA

To further improve the obtained results, we propose combining HOF features with HOG [235] features. OpenCV package [233] is employed to extract HOG features from the developed dataset (SADAM). The first-order gradients are computed from each frame (detected face). These capture contour, silhouette and some texture information, while providing further resistance to illumination variations. Next, the

Table 6.4: Results for multimodal gender, age and sentiment recognition systems when applying HOF and HOG

Modality	Rec	Prc	F_1	GM	Acc	MCC
T-V	98.85	98.91	98.85	99.35	98.85	98.75
A-V	98.85	98.91	98.85	99.35	98.85	98.75
A-T-V	99.24	99.28	99.23	99.56	99.24	99.17

histograms are computed and represented as a feature vector. HOG produces an encoding that is sensitive to local image content while remaining resistant to small changes in pose or appearance. The frame is divided into “cells”. One-dimensional (1-D) histogram of gradient orientations is combined over all the pixels in the cell to form the basic “orientation histogram” representation. Each orientation histogram divides the gradient angle range into a fixed number of predetermined bins. The gradient magnitudes of the pixels in the cell are used to vote into the orientation histogram. The fourth stage computes normalization, which takes local groups of cells and contrasts, normalizes their overall responses before passing to next stage to form HOG descriptors per block. Then, the HOG descriptors from all blocks are combined to represent a feature vector. The average of all feature vectors of frames are calculated to represent each video by one feature vector. The results of multimodal gender, age and sentiment detection system when applying HOG is presented in Table 6.4 and Figure 6.12 compares the results of multimodal gender, age and sentiment detection system before and after applying HOG features.

6.6 Summary

The experiments described in this chapter reveal the following observations::

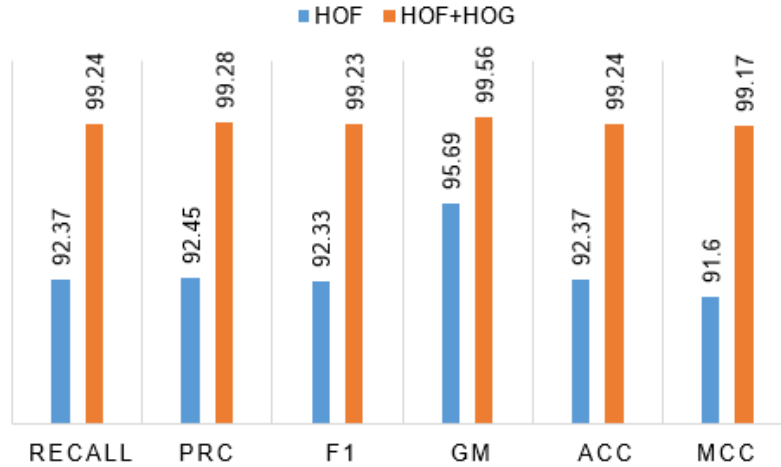


Figure 6.12: Compressions of multimodal gender, age and sentiment recognition at feature level fusion in case of HOF and HOF+HOG visual features

- Multimodal approach improves demographics recognition system significantly comparing to stand-alone systems for gender, age, dialect and nationality.
- Combining different modalities doesn't lead to improving the results only but also minimizing the consumed fitting and scoring time of the visual modality.
- Bi-systems of sentiment and demographics systems are improved in case of multimodal approaches
- Demographics information of users are capable to detect their sentiments. In other words, there is a correlation between demographic characteristics and sentiments
- Incorporating demographics as a new modality improves the results significantly in nearly all cases.
- Further improvements can be achieved when combining other types of features such as HOG.

CHAPTER 7

CONCLUSIONS AND FUTURE DIRECTIONS

With the prevalence of social media and tremendous amount of online data, there is a growing interest in the field of opinion mining and sentiment analysis over the past years. It supports decision making in a wide spectrum of applications including product marketing, customer service, healthcare, politics, etc. Several attempts have been carried out to deal with sentiment analysis tasks. Since modern information systems can process videos and extract interesting patterns much easier than ever before, video opinion mining has recently become one of the active research areas within the machine learning and data mining community.

In this dissertation, different features are adopted and evaluated to textual based sentiment analysis for Arabic, namely *tf-idf* and latent semantic analysis, as hand crafted features, and word embedding based features, as neural language features. Several machine learning classifiers are used to evaluate the proposed features using

different datasets. Word embedding based textual features outperform other textual features in all cases. In addition, traditional features or hand-crafted features have several the limitations including curse of dimensionality, laborious effort to engineering features and high computations. The class imbalance problem of sentiment analysis was not got significant attention in the literature. Some studies balanced their used datasets manually while few studies dealt with class imbalance problem through applying different methods. The class imbalance problem is addressed, in this dissertation, through adopting several oversampling techniques along-side with word embedding techniques with different imbalance ratios. Various deep learning models based on CNN and LSTM for sentiment analysis of Arabic microblogs are investigated. Word2vec are used for vectorizing text and several deep learning architectures using CNN and LSTM are designed and evaluated. The experiments are run on two publicly available Arabic tweets datasets. The highest results have been attained when combining LSTM and compared favorably with most related work.

Moreover, this dissertation conducted a comprehensive analysis for emojis in sentiment analysis. Most related works are reviewed and classified according to the emojis' applications, representations, issues, and approaches. The idea of adopting new non-verbal emoji-based features for sentiment analysis of microblogs is explored. Several types of emojis-based features are proposed and evaluated. We considered 969 emojis and prepared a dataset of 2091 instances expressed in different Arabic dialects such that each instance contains at least one emoji. The suggested features are compared with different textual features using several machine learning algorithms. The experimental results illustrate that emoji-based features alone can be a very effective means

for detecting sentiment polarity with high performance. Additionally, emojis2vec based features especially those generated using skip-gram technique outperform other types of emojis-based features. We observed that users tend to use emojis with positive polarity or happy emotion more than other polarities or emotions. This issue is dealt with as a class imbalance problem and was addressed through generating synthetic instances for the minority class. The proposed method is based on Bootstrap Aggregating (Bagging) algorithm and oversampling methods to build and combine multiple models from the training dataset. Three different classifiers are evaluated as single and ensemble classifiers: naïve Bayes, k -NN and decision trees. The performance is evaluated and compared on varying imbalance ratio ranging from almost two to more than 14. The experimental results show that the proposed approach performs better than the baseline approaches in most of the considered cases.

The first systematic multimodal sentiment analysis for Arabic video opinions is presented in this dissertation. Due to unavailability of Arabic multimodal dataset, significant efforts are made to develop Sentiment Analysis Dataset for Arabic Multimodal (SADAM). Three modalities are considered, namely: text, audio and visual. Features are extracted from each modality and then evaluated individually and in combinations. Different features for each modality are extracted and evaluated. Word embeddings are adopted to represent text modality and a combination of prosodic and spectral features are adopted to detect those characteristics. In addition, dense optical flow technique are adopted to represent visual features. We also evaluated a combination of local and global descriptors. Different fusion methods are presented and evaluated to combine these modalities in different fusion levels: feature level, score

level and decision level. Moreover, multi-level hybrid fusion methods are presented. Different fusion methods are investigated using score fusion rules. Combining different modalities improves the unimodal approaches significantly in all cases. Multi-level fusion method of feature level with score level reports the highest performance.

Multimodal demographics recognition of videos using three modalities is presented and combined with sentiment polarity detection. With the application of same feature extraction, fusion and classification methods, significant improvements are achieved when combining multiple modalities. This encourages us to use demographics as a new modality and evaluate its effectiveness to detect sentiment. Incorporating demographics as a fourth modality with textual, audio and visual modalities to detect sentiment leads to improve the results. Our work is novel and has several applications. For example, it can be integrated in web browsers or as a stand-alone application for accessing social media platforms using computer networks and smartphones. While videos are played, the system can recognize and report sentiment and demographics. This can be very interesting and useful in demographic studies concerning marketing customized products and services, public opinion research on political polls and governmental services, intelligent tutoring systems, etc.

7.1 Contributions

This work delivers the following contributions:

1. Developing the theory of Arabic sentiment analysis in different modalities including text, audio and visual. A comprehensive literature review with explor-

ing and reviewing related works. Taxonomies and frameworks are produced such as a framework for constructing multimodal sentiment analysis datasets and taxonomy for emojis in social media in terms of their applications, issues, representations, and approaches.

2. Empirically evaluating and comparing different techniques to text-based sentiment analysis for Arabic.
 - Employing neural language based features and comparing them with surface features with shallow and deep learning classifiers and investigating several types of textual features.
 - Investigating different sampling techniques to handle the problem of imbalanced distribution of instances among classes.
 - Developing deep learning techniques for Arabic sentiment analysis using CNN, LSTM and hybrid of them.
 - Evaluating the effects of word embedding based features to detect Arabic spam opinions with reporting accepted results.
3. Proposing and evaluating non-verbal (emojis) features for sentiment analysis and then combining them with text modality using different fusion methods in different levels to produce bi-modal sentiment analysis.
4. Multimodal sentiment analysis for Arabic video opinion mining approach.
5. Investigating and proposing different fusion methods to combine them in different fusion levels.

6. Proposing multi-modal, and multi-class system approach to detect demographics of speakers with their sentiments.
7. Presenting a multi-modal, multi-label, multi-class system for sentiment and demographic detection. The developed system can detect sentiment polarity (positive or negative), gender (male or female), four age groups (young-adults, middle-age I, middle-age II and senior) and four language dialects.
8. Conducting intensive experiments to evaluate the proposed methods.
9. Several resources are constructed and developed, including:
 - A dataset for Arabic text based sentiment analysis in which each instance contains at least one emoji.
 - Dataset for multimodal Arabic sentiment analysis with a variety of speakers' ages and gender, nationalities, expressed dialects, recording environments and topics.
 - Prototype systems are developed as a proof of concept.

Our multimodal approach differs from the available approaches in several dimensions including:

- It is the first multimodal approach to deal with Arabic speakers.
- It presents a new direction for sentiment and demographic characteristics detection.

- The used textual features, audio feature and visual features have not been evaluated for multimodal sentiment analysis approaches.
- It performs extensive evaluation for several fusion techniques. Most of the related works just focus on one or two techniques.
- It presents new fusion method based on ensemble neural networks.

7.2 Challenges and Limitations

The unavailability of Arabic resources required us spending a lot of time and paying huge efforts on building our resources. This issue gets more complicated in case of multi-label approach in which some classes has very small number of instances and there is no samples for one case “negative-female-AGA”. There are few available transcription techniques of Arabic language which fail to convert the speech to text especially in case of multi-dialect contents.

Some limitations that need further studies include transcription and segmentation of videos into utterances. A further limitation is that this study just considered positive and negative sentiments. Considering other polarities might affect the performance of the proposed system. Furthermore, although the developed Sentiment Analysis Dataset for Arabic Multimodal (SADAM) dataset is larger than some available multimodal datasets, it needs to be extended in terms of number of videos and sentiment polarities. In addition, the corpus used of building emojis embedding based features is in English with the assumption that emojis are language independent. We proposed performing opinion spam detection as a preprocessing step before evaluating

the opinions and applied them on spam opinion datasets. However, we didn't apply it as preprocessing step before detecting sentiments because there is no textual dataset annotated as spam/no-spam along with positive/negative/neutral to evaluate or validate our claim. Scalability was analyzed and investigated for some tasks including addressing imbalance class problem. However, it needs more analysis for multimodal sentiment and demographic recognition approaches.

7.3 Future Directions

- Extending multimodal Arabic dataset to include multi-sentiment levels.
- Investigating different features to represent different modalities.
- Exploring deep learning architectures with the extended multimodal dataset version.
- Addressing the issue of Arabic transcription.
- Proposing automated techniques for segmenting audio and visual modalities.

APPENDIX A

LIST OF PUBLICATIONS

A.1 Conferences

1. Sadam Al-Azani, El-Sayed M. El-Alfy: Using Word Embedding and Ensemble Learning for Highly Imbalanced Data Sentiment Analysis in Short Arabic Text. ANT/SEIT 2017: 359-366
2. Sadam Al-Azani, El-Sayed M. El-Alfy: Hybrid Deep Learning for Sentiment Polarity Determination of Arabic Microblogs. The International Conference on Neural Information Processing ICONIP (2) 2017: 491-500
3. Sadam Al-Azani, El-Sayed M. El-Alfy: Combining Emojis with Arabic textual features for sentiment classification. In the 9th International Conference on Information and Communication Systems (ICICS), pp. 139-144. IEEE, 2018.
4. Sadam Al-Azani, El-Sayed M. El-Alfy: Emojis-Based Sentiment Classification of Arabic Microblogs Using Deep Recurrent Neural Networks, The Second International Conference on Computing Sciences and Engineering (ICCSE). IEEE,

2018

5. Sadam Al-Azani, El-Sayed M. El-Alfy: Imbalanced Sentiment Polarity Detection Using Emoji-Based Features and Bagging Ensemble, 1st International Conference on Computer Applications & Information Security (ICCAIS 2018)
6. Sadam Al-Azani, El-Sayed M. El-Alfy: Social Sensing Applications and Case Study Using Acoustic Arabic Opinion Mining, Smart Cities Symposium 2018
7. Sadam Al-Azani, El-Sayed M. El-Alfy: Emoji-Based Sentiment Analysis of Arabic Microblogs Using Machine Learning, 21st Saudi Computer Society National Computer Conference (SCS-NCC'2018)
8. El-Sayed M. El-Alfy, Sadam Al-Azani: Statistical Comparison of Opinion Spam Detectors in Social Media with Imbalanced Datasets (SSCC-2018)
9. Sadam Al-Azani, El-Sayed M. El-Alfy: Detection of Arabic Spam Tweets Using Word Embedding and Machine Learning, the 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT'18), IEEE, 2018
10. Sadam Al-Azani, El-Sayed M. El-Alfy: Using Feature-Level Fusion for Multimodal Gender Recognition for Opinion Mining Videos, the 8th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO-2019)
11. Sadam Al-Azani, El-Sayed M. El-Alfy: Multimodal Sentiment and Gender Classification for Video Logs, the 11th International Conference on Agents and Artificial Intelligence (ICCART-2019).

12. Sadam Al-Azani, El-Sayed M. El-Alfy: Audio-Textual Arabic Dialect Identification for Opinion Mining Videos, the IEEE Symposium Series on Computational Intelligence (IEEE SSCI2019)

A.2 Journals

1. Sadam Al-Azani, El-Sayed M. El-Alfy: Multimodal Age Group Recognition for Opinion Video Logs using Ensemble of Neural Networks, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 10, No. 4, 2019, ISI Journal
2. Sadam Al-Azani, El-Sayed M. El-Alfy: Early and Late Level Fusion of Emojis and Text to Enhance Arabic Sentiment Analysis in Social Media. Expert Systems and Applications. **(Submitted)**
3. Empirical Study on Imbalanced Learning of Arabic Sentiment Polarity with Neural Word Embedding **(Under preparation)**
4. Multimodal Arabic Sentiment Analysis for Video Opinion Mining **(Under preparation)**

A.3 Patents

1. El-Sayed M. El-Alfy, Sadam Al-Azani: Multi-modal Detection Engine of Sentiment and Demographic Characteristics for Social Media Videos. U.S. patent KFUPM Reference #: 2018-185 **(Submitted)**

APPENDIX B

STATISTICAL TESTS

A taxonomy to perform a significant test is proposed as shown in Figure B.1. It can be extended to include different cases and types of tests.

McNemar's Test Statistic

Given a training set and a validation set, we train algorithms A and B on the training set to obtain two classifiers \hat{f}_A and \hat{f}_B , respectively and test them using the validation set. McNemar's test "chi-squared" (χ^2) [236] can be used to compare the two classifiers \hat{f}_A and \hat{f}_B in terms of the predictive accuracy.

Given two models, the null and alternative hypotheses can be formulated as follows:

$H_0 : b = c$, $H_1 : b \neq c$ where, b is number of examples misclassified by \hat{f}_A but not by \hat{f}_B , c is number of examples misclassified by \hat{f}_B but not by \hat{f}_A . The χ^2 statistic with one degree of freedom is

$$\chi^2 = \frac{(|b - c|)^2}{b + c} \quad (\text{B.1})$$

such that the value of $b + c$ must be large. p -value can be computed after defining

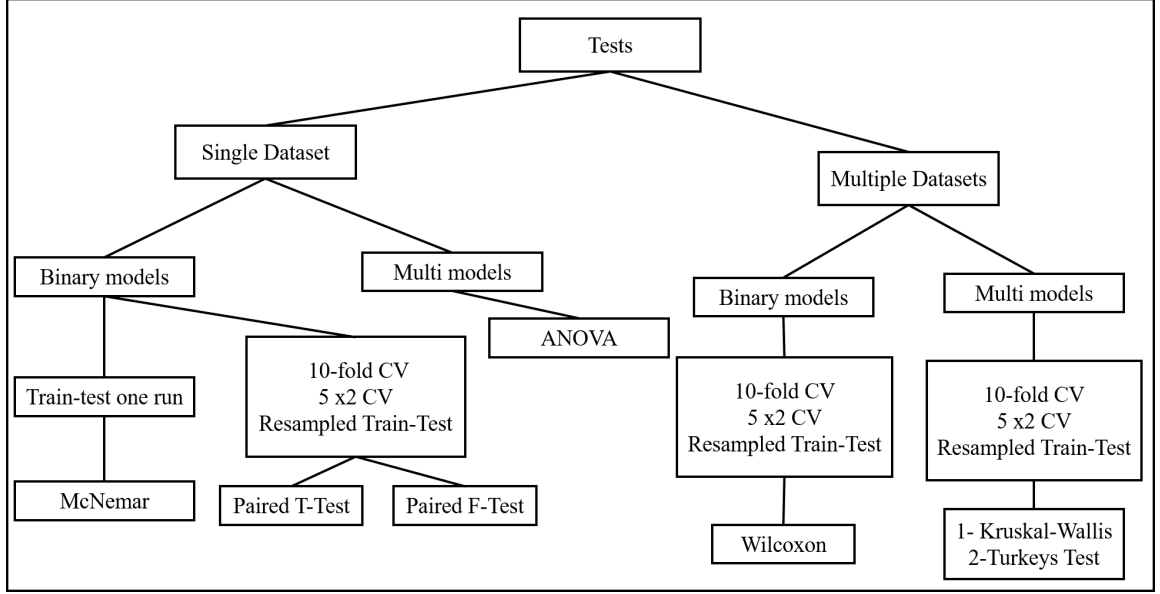


Figure B.1: Significance Tests Taxonomy

a significance threshold, $\alpha = 0.05$. The null hypothesis can be rejected if the *p-value* less than the significance threshold. Continuity correction for χ^2 is proposed by Edward [237] to the fact that χ^2 is continuous while the statistic is discrete by subtracting value of one from the numerator

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \quad (\text{B.2})$$

χ^2 value may not be well-approximated by the χ^2 distribution for small sample sizes (when $b + c < 25$), so exact binomial test is recommended to compute the exact *p-value*.

$$\text{exact-}p\text{-value} = 2 \sum_{i=b}^n \binom{n}{i} 0.5^i (1 - 0.5)^{n-i} \quad (\text{B.3})$$

where $n = b + c$

“McNemar”’s test can be used in the case if we have one training and one validation

set only.” [238]. Therefore, other approaches should be used in case of running several iterations such as : resampled training-testing, k -fold cross-validation or 5×2 cross-validation.

Paired t Test

- k -fold cross-validation paired t test: The dataset is divided, randomly, into k equal disjoint sets (folds), T_1, T_2, \dots, T_k . k -trials is conducted such that in each trail one fold i is used as the test set, and the training set are combination of the remaining folds. For each fold i , let p_i^1 and p_i^2 the error percentages of the classifiers \hat{f}_A and \hat{f}_B on the validation sets, and $p_i = p_i^1 - p_i^2$. The two classification algorithms have the same or equivalent error rate, if the difference of there means is zero; i.e, $p_i = 0$. The null and alternative hypotheses are $H_0 : \mu = 0$, $H_1 : \mu \neq 0$

$$t = \frac{\bar{p} \cdot \sqrt{k}}{\sqrt{\frac{\sum_{i=1}^k (p_i - \bar{p})^2}{k-1}}} \quad (\text{B.4})$$

where $\bar{p} = \frac{1}{n} \sum_{i=1}^n p_i$. Under the null hypothesis ($H_0 : \mu = 0$), this statistic has a t -distribution with $k - 1$ degrees of freedom. At significance level α , the null hypothesis is rejected if this value doesn't belong to the interval $(-t_{\alpha/2, k-1}, t_{\alpha/2, k-1})$.

One-tailed test is used to test whether the first algorithm has less error than the second. The one-sided null and alternative hypotheses are: $H_0 : \mu \geq 0$, $H_1 : \mu < 0$.

- Resampled paired t test: it is similar to k -fold cross-validation where each pair of training and test sets is constructed by randomly dividing the dataset n trails.
- 5×2 -fold cross-validation t test: proposed by Dietterich [239] to overcomes the problem of underestimated variance and the consequently elevated Type I error (which means there is a difference between the tested algorithms while in fact there is not) of the k -fold cross validation. Five replications of two-fold cross-validation are performed.

5×2 -fold cross-validation has acceptable type I error [239, 240]. However, it has not been widely accepted and is not considered as powerful as 10-fold cross validation in data mining community [240].

REFERENCES

- [1] B. Liu, “Sentiment analysis and opinion mining,” *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.
- [2] M. Farhadloo and E. Rolland, “Fundamentals of sentiment analysis and its applications,” in *Sentiment Analysis and Ontology Engineering*. Springer, 2016, pp. 1–24.
- [3] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177.
- [4] 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*. Association for Computational Linguistics, 2005, pp. 579–586.
- [5] H. Yu and V. Hatzivassiloglou, “Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences,” in *Proceedings of the International Conference on Empirical Methods in Natural*

- Language Processing*. Association for Computational Linguistics, 2003, pp. 129–136.
- [6] P. Carvalho, L. Sarmiento, J. Teixeira, and M. J. Silva, “Liars and saviors in a sentiment annotated corpus of comments to political debates,” in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 564–568.
 - [7] L. Lloyd, D. Kechagias, and S. Skiena, “Lydia: A system for large-scale news analysis,” in *Proceedings of the International Symposium on String Processing and Information Retrieval*, 2005, pp. 161–166.
 - [8] G. Carenini, R. T. Ng, and X. Zhou, “Summarizing emails with conversational cohesion and subjectivity,” in *Proceedings of Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, vol. 8, 2008, pp. 353–361.
 - [9] K. Denecke and Y. Deng, “Sentiment analysis in medical settings: New opportunities and challenges,” *Artificial Intelligence in Medicine*, vol. 64, no. 1, pp. 17–27, 2015.
 - [10] K. Ravi and V. Ravi, “A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,” *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.
 - [11] A. Sutcliffe, “Multimedia user interface design,” in *The human-computer interaction handbook*, 2007, pp. 402–420.

- [12] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR)*, vol. 47, no. 3, p. 43, 2015.
- [13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
- [14] Wikipedia, "Arabic," <https://en.wikipedia.org/wiki/Arabic>, 2019, last accessed, February 2019.
- [15] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short Arabic text," *Procedia Computer Science*, vol. 109, pp. 359–366, 2017.
- [16] —, "Hybrid deep learning for sentiment polarity determination of Arabic microblogs," in *Proceedings of the International Conference on Neural Information Processing*, 2017, pp. 491–500.
- [17] —, "Imbalanced sentiment polarity detection using emoji-based features and bagging ensemble," in *Proceedings of the 1st IEEE International Conference on Computer Applications & Information Security (ICCAIS)*, 2018, pp. 1–5.
- [18] —, "Combining emojis with Arabic textual features for sentiment classification," in *Proceedings of the 9th IEEE International Conference on Information and Communication Systems (ICICS)*, 2018, pp. 139–144.

- [19] —, “Emoji-based sentiment analysis of Arabic microblogs using machine learning,” in *Proceedings of the 21st IEEE Saudi Computer Society National Computer Conference (NCC)*, 2018, pp. 1–6.
- [20] —, “Emojis-based sentiment classification of Arabic microblogs using deep recurrent neural networks,” in *Proceedings of the IEEE International Conference on Computing Sciences and Engineering (ICCSE)*, 2018, pp. 1–6.
- [21] —, “Social sensing applications and case study using acoustic arabic opinion mining,” 2018.
- [22] —, “Using feature-level fusion for multimodal gender recognition for opinion mining videos,” in *Proceedings of the 8th IEEE International Conference on Modeling, Simulation and Applied Optimization (ICMSAO)*, 2019.
- [23] —, “Multimodal sentiment and gender classification for video logs,” in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence (ICCART)*, 2019.
- [24] “Multimodal age-group recognition for opinion video logs using ensemble of neural networks,” *International Journal of Advanced Computer Science and Applications*.
- [25] A. Onan, S. Korukoğlu, and H. Bulut, “A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification,” *Expert Systems with Applications*, 2016.

- [26] G. Wang, Z. Zhang, J. Sun, S. Yang, and C. Larson, “POS-RS: A random subspace method for sentiment classification based on part-of-speech analysis,” *Information Processing and Management*, vol. 51, no. 4, pp. 458–479, 2015.
- [27] J. Cotelo, F. Cruz, F. Enríquez, and J. Troyano, “Tweet categorization by combining content and structural knowledge,” *Information Fusion*, vol. 31, pp. 54–64, 2016.
- [28] F. Wu, Y. Song, and Y. Huang, “Microblog sentiment classification with heterogeneous sentiment knowledge,” *Information Sciences*, vol. 373, pp. 149–164, 2016.
- [29] S. Rill, D. Reinel, J. Scheidt, and R. V. Zicari, “Politwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis,” *Knowledge-Based Systems*, vol. 69, pp. 24–33, 2014.
- [30] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, “Sentiment, emotion, purpose, and style in electoral tweets,” *Information Processing & Management*, vol. 51, no. 4, pp. 480–499, 2015.
- [31] M. Abdul-Mageed, M. Diab, and S. Kübler, “SAMAR: Subjectivity and sentiment analysis for Arabic social media,” *Computer Speech & Language*, vol. 28, no. 1, pp. 20–37, 2014.
- [32] R. Duwairi, N. A. Ahmed, and S. Y. Al-Rifai, “Detecting sentiment embedded in Arabic social media—a lexicon-based approach,” *Journal of Intelligent & Fuzzy Systems*, vol. 29, no. 1, pp. 107–117, 2015.

- [33] B. Al Shboul, M. Al-Ayyoub, and Y. Jararweh, "Multi-way sentiment classification of Arabic reviews," in *Proceedings of the 6th IEEE International Conference on Information and Communication Systems (ICICS)*, 2015, pp. 206–211.
- [34] R. T. Khasawneh, H. A. Wahsheh, I. M. Alsmadi, and M. N. Al-Kabi, "Arabic sentiment polarity identification using a hybrid approach," in *Proceedings of the 6th IEEE International Conference on Information and Communication Systems (ICICS)*, 2015, pp. 148–153.
- [35] B. Brahimi, M. Touahria, and A. Tari, "Data and text mining techniques for classifying Arabic tweet polarity," *Journal of Digital Information Management*, vol. 14, no. 1, p. 15, 2016.
- [36] N. Omar, M. Albared, T. Al-Moslmi, and A. Al-Shabi, "A comparative study of feature selection and machine learning algorithms for Arabic sentiment classification," in *Asia Information Retrieval Symposium*. Springer, 2014, pp. 429–443.
- [37] A. M. Rabab'ah, M. Al-Ayyoub, Y. Jararweh, and M. N. Al-Kabi, "Evaluating sentistrength for Arabic sentiment analysis," in *Proceedings of the 7th IEEE International Conference on Computer Science and Information Technology (CSIT)*, 2016, pp. 1–6.
- [38] H. ElSahar and S. R. El-Beltagy, "Building large Arabic multi-domain resources for sentiment analysis," in *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2015, pp. 23–34.

- [39] A. A. Al Sallab, R. Baly, G. Badaro, H. Hajj, W. El Hajj, and K. B. Shaban, "Deep learning models for sentiment analysis in Arabic," in *ANLP Workshop*, 2015, p. 9.
- [40] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [41] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theories of Emotion*, vol. 1, pp. 3–31, 1980.
- [42] S. Hiai and K. Shimada, "A sarcasm extraction method based on patterns of evaluation expressions," in *Proceedings of the 5th International IEEE Congress on Advanced Applied Informatics (IIAI-AAI)*, 2016, pp. 31–36.
- [43] H. Xu, F. Zhang, and W. Wang, "Implicit feature identification in chinese reviews using explicit topic mining model," *Knowledge-Based Systems*, vol. 76, pp. 166–175, 2015.
- [44] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature based approach for sentiment analysis by using support vector machine," in *Proceedings of the IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 3–8.
- [45] M. Z. Asghar, S. Ahmad, M. Qasim, S. R. Zahra, and F. M. Kundi, "Sentihealth: creating health-related sentiment lexicon using hybrid approach," *SpringerPlus*, vol. 5, no. 1, p. 1139, 2016.

- [46] K. Al-Rowaily, M. Abulaish, N. A.-H. Haldar, and M. Al-Rubaian, “Bisal—a bilingual sentiment analysis lexicon to analyze dark web forums for cyber security,” *Digital Investigation*, vol. 14, pp. 53–62, 2015.
- [47] M. N. Al-Kabi, M. A. Al-Ayyoub, I. M. Alsmadi, and H. A. Wahsheh, “A prototype for a standard Arabic sentiment analysis corpus,” *International Arab Journal of Information Technology (IAJIT)*, vol. 13, 2016.
- [48] E. Refaee and V. Rieser, “An Arabic twitter corpus for subjectivity and sentiment analysis,” in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 2268–2273.
- [49] M. Nabil, M. Aly, and A. F. Atiya, “Astd: Arabic sentiment tweets dataset,” in *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519.
- [50] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, “Arabic sentiment analysis: Lexicon-based and corpus-based,” in *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 2013, pp. 1–6.
- [51] S. Mohammad, M. Salameh, and S. Kiritchenko, “How translation alters sentiment,” *Journal of Artificial Intelligence Research*, vol. 55, pp. 95–130, 2016.
- [52] E. Refaee and V. Rieser, “iLab-Edinburgh at SemEval-2016 Task 7: A hybrid approach for determining sentiment intensity of Arabic twitter phrases,” in *Pro-*

- ceedings of the 10th International Workshop on Semantic Evaluation SemEval*, ser. SemEval'16, San Diego, California, June 2016.
- [53] T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 347–354.
 - [54] S. Krishnamoorthy, “Linguistic features for review helpfulness prediction,” *Expert Systems with Applications*, vol. 42, no. 7, pp. 3751–3759, 2015.
 - [55] I. Mani and I. Zhang, “kNN approach to unbalanced data distributions: a case study involving information extraction,” in *Proceedings of workshop on learning from imbalanced datasets*, 2003.
 - [56] S.-J. Yen and Y.-S. Lee, “Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset,” in *Intelligent Control and Automation*, 2006, pp. 731–740.
 - [57] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, “Classification with class imbalance problem: a review,” *International Journal of Advances in Soft Computing and its Application*, vol. 7, no. 3, 2015.
 - [58] T. G. Dietterich, “Ensemble methods in machine learning,” in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
 - [59] A. Hassan, A. Abbasi, and D. Zeng, “Twitter sentiment analysis: A bootstrap

- ensemble framework,” in *Proceedings of the International Conference on Social Computing (SocialCom)*. IEEE, 2013, pp. 357–364.
- [60] J. Ah-Pine and E.-P. Soriano-Morales, “A study of synthetic oversampling for twitter imbalanced sentiment analysis,” in *Workshop on Interactions between Data Mining and Natural Language Processing (DMNLP 2016)*, 2016.
- [61] V. Gopalakrishnan and C. Ramaswamy, “Sentiment learning from imbalanced dataset: an ensemble based method,” *Int. J. Artif. Intell*, vol. 12, no. 2, pp. 75–87, 2014.
- [62] H. Al Najada and X. Zhu, “isrd: Spam review detection with imbalanced data distributions,” in *Proceedings of the IEEE 15th International Conference on Information Reuse and Integration (IRI)*, 2014, pp. 553–560.
- [63] S. Li, Z. Wang, G. Zhou, and S. Y. M. Lee, “Semi-supervised learning for imbalanced sentiment classification,” in *Proceedings International Joint Conference on Artificial Intelligence (IJCAI)*, vol. 22, no. 3, 2011, p. 1826.
- [64] A. Mountassir, H. Benbrahim, and I. Berrada, “Addressing the problem of unbalanced data sets in sentiment analysis,” in *Knowledge Discovery and Information Retrieval (KDIR)*, 2012, pp. 306–311.
- [65] E. Refaee, “Sentiment analysis for micro-blogging platforms in Arabic,” Ph.D. dissertation, Heriot-Watt University, 2016.
- [66] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word

- representations in vector space,” in *Proceedings of Workshop at International Conference on Learning Representations*, 2013.
- [67] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.
- [68] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.
- [69] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, “A convolutional neural network for modelling sentences,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.
- [70] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, “Learning semantic representations using convolutional neural networks for web search,” in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 373–374.
- [71] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [72] G. Liu, X. Xu, B. Deng, S. Chen, and L. Li, “A hybrid method for bilingual text sentiment classification based on deep learning,” in *Proceedings of the 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2016, pp. 93–98.

- [73] A. M. Alayba, V. Palade, M. England, and R. Iqbal, “Arabic language sentiment analysis on health services,” *arXiv preprint arXiv:1702.03197*, 2017.
- [74] M. Abbes, Z. Kechaou, and A. M. Alimi, “Enhanced deep learning models for sentiment analysis in arab social media,” in *Proceedings of the International Conference on Neural Information Processing*, 2017, pp. 667–676.
- [75] T. Mikolov, S. Kombrink, L. Burget, J. Černocký, and S. Khudanpur, “Extensions of recurrent neural network language model,” in *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5528–5531.
- [76] J. Pennington, R. Socher, and C. Manning, “Glove: Global vectors for word representation,” in *Conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [77] A. A. Aziz and L. Tao, “Word embeddings for Arabic sentiment analysis,” in *Proceedings of the IEEE International Conference on Big Data*, vol. 7, 2016, pp. 3820–3825.
- [78] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, “Word embeddings and convolutional neural network for Arabic sentiment classification,” in *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, 2016, pp. 2418–2427.
- [79] N. Boudad, R. Faizi, R. O. H. Thami, and R. Chiheb, “Sentiment analysis in Arabic: A review of the literature,” *Ain Shams Engineering Journal*, 2017.

- [80] C. Grier, K. Thomas, V. Paxson, and M. Zhang, “@spam: the underground on 140 characters or less,” in *Proceedings of the 17th ACM conference on Computer and Communications Security*, 2010, pp. 27–37.
- [81] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” *arXiv preprint arXiv:1703.03107*, 2017.
- [82] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Detecting automation of twitter accounts: Are you a human, bot, or cyborg?” *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 6, pp. 811–824, 2012.
- [83] H. Almerexhi and T. Elsayed, “Detecting automatically-generated Arabic tweets,” in *Asia Information Retrieval Symposium*, 2015, pp. 123–134.
- [84] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, “Detecting spammers on twitter,” in *Proceedings of the Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.
- [85] N. El-Mawass and S. Alaboodi, “Detecting Arabic spammers and content polluters on twitter,” in *Proceedings of the Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*, 2016, pp. 53–58.
- [86] M. Rajdev and K. Lee, “Fake and spam messages: Detecting misinformation during natural disasters on social media,” in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 1, 2015, pp. 17–20.

- [87] T. C. Alberto, J. V. Lochter, and T. A. Almeida, “Tubespam: Comment spam filtering on youtube,” in *Proceedings of the 14th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2015, pp. 138–143.
- [88] A. H. Wang, “Detecting spam bots in online social networking sites: a machine learning approach,” in *IFIP Annual Conference on Data and Applications Security and Privacy*, 2010, pp. 335–342.
- [89] D. Wang, D. Irani, and C. Pu, “A social-spam detection framework,” in *Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*. ACM, 2011, pp. 46–54.
- [90] C. Chen, J. Zhang, X. Chen, Y. Xiang, and W. Zhou, “6 million spam tweets: A large ground truth for timely twitter spam detection,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, 2015, pp. 7065–7070.
- [91] M. Mataoui, O. Zelmati, D. Boughaci, M. Chaouche, and F. Lagoug, “A proposed spam detection approach for Arabic social networks content,” in *Proceedings of the IEEE International Conference on Mathematics and Information Technology (ICMIT)*, 2017, pp. 222–226.
- [92] E.-S. M. El-Alfy and S. Al-Azani, “Statistical comparison of opinion spam detectors in social media with imbalanced datasets,” in *International Symposium on Security in Computing and Communication*. Springer, 2018, pp. 157–167.
- [93] S. Al-Azani and E.-S. M. El-Alfy, “Detection of Arabic spam tweets using word embedding and machine learning,” in *Proceeding of the International IEEE*

Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT'18), 2018.

- [94] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in *Proceedings of the 13th ACM International Conference on Multimodal Interfaces*, 2011, pp. 169–176.
- [95] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, “Youtube movie reviews: Sentiment analysis in an audio-visual context,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.
- [96] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, “Utterance-level multimodal sentiment analysis,” in *Association for Computational Linguistics (ACL)*, 2013, pp. 973–982.
- [97] V. Perez Rosas, R. Mihalcea, and L.-P. Morency, “Multimodal sentiment analysis of spanish online videos,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.
- [98] S. Poria, E. Cambria, and A. Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015, pp. 2539–2544.
- [99] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputing*, vol. 174, pp. 50–59, 2016.

- [100] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Association for Computational Linguistics*, 2017.
- [101] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency, “Multi-level multiple attentions for contextual multimodal sentiment analysis,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 1033–1038.
- [102] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 1114–1125.
- [103] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, “Ensemble of SVM trees for multimodal emotion recognition,” in *Proceedings of the IEEE Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), Asia-Pacific*, 2012, pp. 1–4.
- [104] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, “Towards an intelligent framework for multimodal affective data analysis,” *Neural Networks*, vol. 63, pp. 104–116, 2015.
- [105] M. Pereira, F. Pádua, A. Pereira, F. Benevenuto, and D. Dalip, “Fusing audio, textual, and visual features for sentiment analysis of news videos,” in *Proceedings*

- of the 10th International Conference on Web and Social Media, ICWSM, 2016, pp. 659–662.
- [106] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional MKL based multimodal emotion recognition and sentiment analysis,” in *Proceedings of the IEEE International Conference on Data Mining (ICDM), Barcelona*, 2016.
 - [107] F. Chen, Y. Gao, D. Cao, and R. Ji, “Multimodal hypergraph learning for microblog sentiment prediction,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2015, pp. 1–6.
 - [108] C. Baecchi, T. Uricchio, M. Bertini, and A. Del Bimbo, “A multimodal feature learning approach for sentiment analysis of social network multimedia,” *Multimedia Tools and Applications*, vol. 75, no. 5, pp. 2507–2525, 2016.
 - [109] Y. Yu, H. Lin, J. Meng, and Z. Zhao, “Visual and textual sentiment analysis of a microblog using deep convolutional neural networks,” *Algorithms*, vol. 9, no. 2, p. 41, 2016.
 - [110] Q. You, J. Luo, H. Jin, and J. Yang, “Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia,” in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, pp. 13–22.
 - [111] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages,” *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

- [112] K. Kang, C. Yoon, and E. Y. Kim, “Identifying depressive users in twitter using multimodal analysis,” in *Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2016, pp. 231–238.
- [113] G. Cai and B. Xia, “Convolutional neural networks for multimedia sentiment analysis,” in *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 2015, pp. 159–167.
- [114] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013, pp. 223–232.
- [115] S. Govindaraj and K. Gopalakrishnan, “Intensified sentiment analysis of customer product reviews using acoustic and textual features,” *ETRI Journal*, vol. 38, no. 3, pp. 494–501, 2016.
- [116] C.-H. Wu and W.-B. Liang, “Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels,” *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.
- [117] H. Abburi, E. S. A. Akkireddy, S. V. Gangashetty, and R. Mamidi, “Multimodal sentiment analysis of telugu songs,” in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, 2016, p. 48.
- [118] Y. Wang and L. Guan, “Recognizing human emotional state from audiovisual signals,” *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

- [119] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition,” *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.
- [120] Z. Xie and L. Guan, “Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2013, pp. 1–6.
- [121] A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” CS224N Project, Stanford University, Tech. Rep., 2009.
- [122] C. J. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, 2014.
- [123] A. S. Alqarafi, A. Adeel, M. Gogate, K. Dashitpour, A. Hussain, and T. Durani, “Toward’s arabic multi-modal sentiment analysis,” in *Proceedings of the International Conference in Communications, Signal Processing, and Systems*. Springer, 2017, pp. 2378–2386.
- [124] A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,” *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 3, p. 12, 2008.
- [125] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The enterface’05 audio-visual emotion database,” in *Proceedings of the 22nd IEEE International Conference on Data Engineering Workshops (ICDEW’06)*, 2006, pp. 8–8.

- [126] H. Najadat and F. Abushaqra, “Multimodal sentiment analysis of arabic videos,” *Journal of Image and Graphics*, vol. 6, no. 1, 2018.
- [127] B. Schuller, J. Schenk, G. Rigoll, and T. Knaup, ““the godfather” vs.“chaos”: Comparing linguistic analysis based on on-line knowledge sources and bags-of-n-grams for movie review valence estimation,” in *Proceedings of the 10th IEEE International Conference on Document Analysis and Recognition*, 2009, pp. 858–862.
- [128] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, “The computer expression recognition toolbox (cert),” in *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*, 2011, pp. 298–305.
- [129] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.
- [130] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, “The semaine corpus of emotionally coloured character interactions,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2010, pp. 1079–1084.
- [131] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, “The Kaldi speech recognition toolkit,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

- [132] F. Eyben, M. Wöllmer, and B. Schuller, “OpenEAR-introducing the munich open-source emotion and affect recognition toolkit,” in *Proceedings of the 3rd IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–6.
- [133] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
- [134] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, “Microblog sentiment analysis based on cross-media bag-of-words model,” in *Proceedings of International Conference on Internet Multimedia Computing and Service*. ACM, 2014, p. 76.
- [135] Q. You, J. Luo, H. Jin, and J. Yang, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” *arXiv preprint arXiv:1509.06041*, 2015.
- [136] —, “Robust image sentiment analysis using progressively trained and domain transferred deep networks,” in *Proceedings of the National Conference on Artificial Intelligence*, 2015, pp. 381–388.
- [137] L. I. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004.
- [138] S. Robertson, “Understanding inverse document frequency: on theoretical arguments for IDF,” *Journal of Documentation*, vol. 60, no. 5, pp. 503–520, 2004.

- [139] S. T. Dumais, “Latent semantic analysis,” *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, 2004.
- [140] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Processes*, vol. 25, no. 2-3, pp. 259–284, 1998.
- [141] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.
- [142] C. C. Aggarwal, *Data classification: algorithms and applications*. Chapman and Hall/CRC Press, 2014.
- [143] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [144] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of Statistics*, pp. 1189–1232, 2001.
- [145] ———, “Stochastic gradient boosting,” *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.
- [146] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.
- [147] D. Scherer, A. Müller, and S. Behnke, “Evaluation of pooling operations in convolutional architectures for object recognition,” in *Proceedings of the International Conference on Artificial Neural Networks*. Springer, 2010, pp. 92–101.
- [148] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov,

- “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [149] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [150] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [151] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [152] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, “A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2012.
- [153] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, pp. 321–357, 2002.
- [154] H. Han, W.-Y. Wang, and B.-H. Mao, “Borderline-SMOTE: a new over-sampling

- method in imbalanced data sets learning,” in *Proceedings of the International Conference on Intelligent Computing*. Springer, 2005, pp. 878–887.
- [155] H. M. Nguyen, E. W. Cooper, and K. Kamei, “Borderline over-sampling for imbalanced data classification,” *International Journal of Knowledge Engineering and Soft Data Paradigms*, vol. 3, no. 1, pp. 4–21, 2011.
- [156] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2008, pp. 1322–1328.
- [157] M. Kubat, S. Matwin *et al.*, “Addressing the curse of imbalanced training sets: one-sided selection,” in *ICML*, vol. 97. Nashville, USA, 1997, pp. 179–186.
- [158] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, “An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics,” *Information Sciences*, vol. 250, pp. 113–141, 2013.
- [159] V. Ganganwar, “An overview of classification algorithms for imbalanced datasets,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 4, pp. 42–47, 2012.
- [160] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

- [161] G. Lemaître, F. Nogueira, and C. K. Aridas, “Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning,” *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.
- [162] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [163] J. F. Asteroff, “Paralanguage in electronic mail: A case study,” Ph.D. dissertation, Columbia University, New York, 1987.
- [164] D. W. Sanderson and D. Dougherty, *Smileys*. O’Reilly Media, 1993.
- [165] S. Spina, “Role of emoticons as structural markers in twitter interactions,” *Discourse Processes*, pp. 1–18, 2018.
- [166] G. Guibon, M. Ochs, and P. Bellot, “From emojis to sentiment analysis,” in *Workshop Affect Compagnon Artificiel Interaction (WACAI 2016)*, 2016.
- [167] A. Mourad and K. Darwish, “Subjectivity and sentiment analysis of modern standard Arabic and Arabic microblogs,” in *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2013, pp. 55–64.
- [168] E. Refaee and V. Rieser, “Can we read emotions from a smiley face? emoticon-based distant supervision for subjectivity and sentiment analysis of Arabic twitter feeds,” in *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, LREC*, 2014.

- [169] C. Tauch and E. Kanjo, “The roles of emojis in mobile phone notifications,” in *Proceedings of ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 1560–1565.
- [170] F. Barbieri, F. Ronzano, and H. Saggion, “What does this emoji mean? a vector space skip-gram model for twitter emojis,” in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, 2016, pp. 3967–3972.
- [171] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, “A semantics-based measure of emoji similarity,” *arXiv preprint arXiv:1707.04653*, 2017.
- [172] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, “Sentiment of emojis,” *PloS one*, vol. 10, no. 12, p. e0144296, 2015.
- [173] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. De Jong, and U. Kaymak, “Exploiting emoticons in polarity classification of text,” *Journal of Web Engineering*, vol. 14, no. 1&2, pp. 22–40, 2015.
- [174] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, vol. 10, no. 2010, 2010, pp. 1320–1326.
- [175] K.-L. Liu, W.-J. Li, and M. Guo, “Emoticon smoothed language models for twitter sentiment analysis,” in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 2012.

- [176] J. Read, “Using emoticons to reduce dependency in machine learning techniques for sentiment classification,” in *Proceedings of the ACL student research workshop*. Association for Computational Linguistics, 2005, pp. 43–48.
- [177] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, “Arasenti-tweet: A corpus for Arabic sentiment analysis of saudi tweets,” *Procedia Computer Science*, vol. 117, pp. 63–72, 2017.
- [178] W. A. Hussien, Y. M. Tashtoush, M. Al-Ayyoub, and M. N. Al-Kabi, “Are emoticons good enough to train emotion classifiers of Arabic tweets?” in *Proceedings of the 7th IEEE International Conference on Computer Science and Information Technology (CSIT)*, 2016, pp. 1–6.
- [179] F. Barbieri, M. Ballesteros, and H. Saggion, “Are emojis predictable?” *arXiv preprint arXiv:1702.07285*, 2017.
- [180] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. E. Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, “Semeval 2018 task 2: Multilingual emoji prediction,” in *Proceedings of the 12th International Workshop on Semantic Evaluation*, 2018, pp. 24–33.
- [181] F. Barbieri, M. Ballesteros, F. Ronzano, and H. Saggion, “Multimodal emoji prediction,” in *In Proceedings of NAACL: Short Papers, New Orleans, US. Association for Computational Linguistics*, 2018.
- [182] M. Fernández-Gavilanes, J. Juncal-Martínez, S. García-Méndez, E. Costa-Montenegro, and F. J. González-Castaño, “Creating emoji lexica from unsuper-

- vised sentiment analysis of their descriptions,” *Expert Systems with Applications*, vol. 103, pp. 74–91, 2018.
- [183] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, “Blissfully happy” or “ready to fight”: Varying interpretations of emoji,” *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, vol. 2016, 2016.
- [184] H. J. Miller, D. Kluver, J. Thebault-Spieker, L. G. Terveen, and B. J. Hecht, “Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication,” in *Proceedings of the 11th International Conference on Weblogs and Social Media (ICWSM)*, 2017, pp. 152–161.
- [185] S. Annamalai and S. N. A. Salam, “Undergraduates’ interpretation on whatsapp smiley emoji,” *Jurnal Komunikasi, Malaysian Journal of Communication*, vol. 33, no. 4, 2017.
- [186] A. Wolf, “Emotional expression online: Gender differences in emoticon use,” *CyberPsychology & Behavior*, vol. 3, no. 5, pp. 827–833, 2000.
- [187] C. C. Tossell, P. Kortum, C. Shepard, L. H. Barg-Walkow, A. Rahmati, and L. Zhong, “A longitudinal study of emoticon use in text messaging from smartphones,” *Computers in Human Behavior*, vol. 28, no. 2, pp. 659–663, 2012.
- [188] Z. Chen, X. Lu, S. Shen, W. Ai, X. Liu, and Q. Mei, “Through a gender lens: An empirical study of emoji usage over large-scale android users,” *arXiv preprint arXiv:1705.05546*, 2017.

- [189] N. Ljubešić and D. Fišer, “A global analysis of emoji usage,” in *Proceedings of the 10th Web as Corpus Workshop*, 2016, pp. 82–89.
- [190] F. Barbieri, G. Kruszewski, F. Ronzano, and H. Saggion, “How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics,” in *Proceedings of the ACM on Multimedia Conference*, 2016, pp. 531–535.
- [191] X. Lu, W. Ai, X. Liu, Q. Li, N. Wang, G. Huang, and Q. Mei, “Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users,” in *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 770–780.
- [192] G. W. Tigwell and D. R. Flatla, “Oh that’s what you meant!: reducing emoji misunderstanding,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct.* ACM, 2016, pp. 859–866.
- [193] F. Morstatter, K. Shu, S. Wang, and H. Liu, “Cross-platform emoji interpretation: Analysis, a solution, and applications,” *arXiv preprint arXiv:1709.04969*, 2017.
- [194] “Full emoji list, v11.0,” <https://unicode.org/emoji/charts/full-emoji-list.html>, 2019, last accessed, Feb 2019.
- [195] J. Berengueres and D. Castro, “Differences in emoji sentiment perception be-

- tween readers and writers,” in *Proceedings of the IEEE International Conference on Big Data*, 2017, pp. 4321–4328.
- [196] H. Cui, Y. Lin, and T. Utsuro, “Sentiment analysis of tweets by CNN utilizing tweets with emoji as training data,” in *Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM’18)*, 2018.
- [197] L. K. Kaye, S. A. Malone, and H. J. Wall, “Emojis: Insights, affordances, and possibilities for psychological science,” *Trends in cognitive sciences*, vol. 21, no. 2, pp. 66–68, 2017.
- [198] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, “emoji2vec: Learning emoji representations from their description,” *arXiv preprint arXiv:1609.08359*, 2016.
- [199] Y. Tian, T. Galery, G. Dulcinati, E. Molimpakis, and C. Sun, “Facebook sentiment: Reactions and emojis,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017, pp. 11–16.
- [200] T. Tran, D. Nguyen, A. Nguyen, and E. Golen, “Sentiment analysis of marijuana content via facebook emoji-based reactions,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, 2018, pp. 1–6.
- [201] J. B. Walther and K. P. D’Addario, “The impacts of emoticons on message interpretation in computer-mediated communication,” *Social science computer review*, vol. 19, no. 3, pp. 324–347, 2001.

- [202] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran, “Emojinet: Building a machine readable sense inventory for emoji,” in *Proceedings of the International Conference on Social Informatics*, 2016, pp. 527–541.
- [203] —, “Emojinet: An open service and api for emoji sense discovery,” *arXiv preprint arXiv:1707.04652*, 2017.
- [204] N. Na’aman, H. Provenza, and O. Montoya, “Mojisem: Varying linguistic purposes of emoji in (twitter) context,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Student Research Workshop*, 2017, pp. 136–141.
- [205] U. Pavalanathan and J. Eisenstein, “Emoticons vs. emojis on twitter: A causal inference approach,” *AAAI Spring Symposium on Observational Studies through Social Media and Other Human-Generated Content*, 2016.
- [206] H. Cramer, P. de Juan, and J. Tetreault, “Sender-intended functions of emojis in us messaging,” in *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2016, pp. 504–509.
- [207] S. R. El-Beltagy, A. B. Soliman *et al.*, “Niletmrg at semeval-2017 task 4: Arabic sentiment analysis,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017, pp. 790–795.
- [208] R. Baly, G. Badaro, A. Hamdi, R. Moukalled, R. Aoun, G. El-Khoury, A. Al Sallab, H. Hajj, N. Habash, K. Shaban *et al.*, “Omam at semeval-2017 task 4:

- Evaluation of english state-of-the-art sentiment analysis models for Arabic and a new topic-based model,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017, pp. 603–610.
- [209] R. M. Duwairi, R. Marji, N. Sha’ban, and S. Rushaidat, “Sentiment analysis in Arabic tweets,” in *Proceedings of the 5th IEEE International Conference on Information and communication systems (ICICS)*, 2014, pp. 1–6.
- [210] M. El-Masri, N. Altrabsheh, H. Mansour, and A. Ramsay, “A web-based tool for Arabic sentiment analysis,” *Procedia Computer Science*, vol. 117, pp. 38–45, 2017.
- [211] S. Alowaidi, M. Saleh, and O. Abulnaja, “Semantic sentiment analysis of Arabic texts,” *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, pp. 256–262, 2017.
- [212] R. Meo and E. Sulis, “Processing affect in social media: a comparison of methods to distinguish emotions in tweets,” *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 1, p. 7, 2017.
- [213] S. Rosenthal, N. Farra, and P. Nakov, “Semeval-2017 task 4: Sentiment analysis in twitter,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*, 2017, pp. 502–518.
- [214] K. Kira and L. A. Rendell, “A practical approach to feature selection,” in *Proceedings of the 9th International Workshop on Machine Learning*, 1992, pp. 249–256.

- [215] J. Yao, “emoji2vec,” <https://github.com/jiali-ms/emoji2vec/tree/master/data>, 2018, last accessed, Feb 2019.
- [216] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, “Aravec: A set of Arabic word embedding models for use in Arabic nlp,” in *Proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing)*, vol. 117, 2017, pp. 256–265.
- [217] N. Y. Habash, “Introduction to arabic natural language processing,” *Synthesis Lectures on Human Language Technologies*, vol. 3, no. 1, pp. 1–187, 2010.
- [218] O. F. Zaidan and C. Callison-Burch, “Arabic dialect identification,” *Computational Linguistics*, vol. 40, no. 1, pp. 171–202, 2014.
- [219] K. R. Anne, S. Kuchibhotla, and H. D. Vankayalapati, *Acoustic modeling for emotion recognition*. Springer, 2015.
- [220] S. Kuchibhotla, H. Vankayalapati, R. Vaddi, and K. R. Anne, “A comparative analysis of classifiers in emotion recognition through acoustic features,” *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, 2014.
- [221] D. Fortun, P. Bouthemy, and C. Kervrann, “Optical flow modeling and computation: a survey,” *Computer Vision and Image Understanding*, vol. 134, pp. 1–21, 2015.
- [222] K. Mase and A. Pentland, “Automatic lipreading by optical-flow analysis,” *Systems and Computers in Japan*, vol. 22, no. 6, pp. 67–76, 1991.

- [223] C. J. Duthoit, T. Sztynka, S. K. Lal, B. T. Jap, and J. I. Agbinya, “Optical flow image analysis of facial expressions of human emotion: forensic applications,” in *Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop*, 2008, p. 5.
- [224] B. D. Lucas, T. Kanade *et al.*, “An iterative image registration technique with an application to stereo vision,” in *Proceedings of Imaging Understanding Workshop*, 1981, pp. 121–130.
- [225] J. Shi and C. Tomasi, “Good features to track,” Cornell University, Tech. Rep., 1993.
- [226] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.
- [227] P. Viola and M. J. Jones, “Robust real-time face detection,” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [228] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, pp. I–I.
- [229] M. S. Ryoo and L. Matthies, “First-person activity recognition: What are they doing to me?” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2730–2737.

- [230] A. Solichin, A. Harjoko, and A. E. Putra, “Movement direction estimation on video using optical flow analysis on multiple frames,” *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 9, no. 6, pp. 174–181, 2018.
- [231] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.
- [232] T. Giannakopoulos, “pyaudioanalysis: An open-source python library for audio signal analysis,” *PloS One*, vol. 10, no. 12, 2015.
- [233] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library*. O’Reilly Media, Inc., 2008.
- [234] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, “Orb: An efficient alternative to sift or surf,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2564–2571.
- [235] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
- [236] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [237] A. L. Edwards, “Note on the “correction for continuity” in testing the significance of the difference between correlated proportions,” *Psychometrika*, vol. 13, no. 3, pp. 185–187, 1948.

- [238] M. Wozniak, *Hybrid classifiers: methods of data, knowledge, and classifier combination*. Springer, 2013, vol. 519.
- [239] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural Computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [240] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-validation,” in *Encyclopedia of database systems*. Springer, 2009, pp. 532–538.

VITAE

- Name: Saddam Hussein Mohammed Al-Azani
- Nationality: Yemeni
- Date of Birth: 02/01/1982
- Email: *sadamhalazani@gmail.com*
- Permanent Address: Yemen

Academic Background: Saddam Alazani received Bachelor of Science (B.S.) in Computer Science from Thamar University, Yemen in 2004 with honors. He joined the Faculty of Computer Sciences and Information Systems as full time lecture at Thamar University, Thamar, Yemen, from 2007 to 2010. He joined the Information and Computer science department as full time student at King Fahd University of Petroleum & Minerals (KFUPM), Dhahran, Saudi Arabia in September 2011 and obtained Master of Science (M.S.) in Computer science from KFUPM in April 2014, with a GPA of 3.75/4.0. Al-Azani got the opportunity to pursue his PhD studies in KFUPM as a full time student. He completed his PhD in Computer Science in May 2019. His research

interests are related to Machine Learning, Natural Language Processing, Computer Vision, Pattern Recognition, Social Network Big Data Analytics, Deep Learning. Al-Azani has several publications in various international journals and conferences. He has served as a technical program committee member and a reviewer of international conferences and journals.