

بسم الله الرحمن الرحيم



HOSSAM MAGHRABY



شبكة المعلومات الجامعية التوثيق الالكتروني والميكرو فيلم



HOSSAM MAGHRABY

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
على هذه الأقراص المدمجة قد أعدت دون أية تغيرات



يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار



HOSSAM MAGHRABY



HILATSA: A HYBRID INCREMENTAL LEARNING APPROACH FOR ARABIC TWEETS SENTIMENT ANALYSIS

By

Kariman Mahmoud Hamouda Elshakankery

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

HILATSA: A HYBRID INCREMENTAL LEARNING APPROACH FOR ARABIC TWEETS SENTIMENT ANALYSIS

By
Kariman Mahmoud Hamouda Elshakankery

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Under the Supervision of

**Prof. Dr. Magda Bahaa Eldin
Fayek**

Professor
Computer Engineering
Faculty of Engineering, Cairo University

Dr. Mona Farouk Ahmed

Associate Professor
Computer Engineering
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

HILATSA: A HYBRID INCREMENTAL LEARNING APPROACH FOR ARABIC TWEETS SENTIMENT ANALYSIS

By
Kariman Mahmoud Hamouda Elshakankery

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
in
Computer Engineering

Approved by

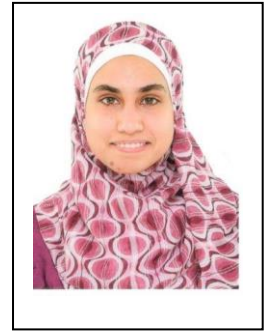
Prof. Dr. Magda Bahaa ElDin Fayek, Thesis Main Advisor

Prof. Dr. Nevin Mahmoud Darwish, Internal Examiner

**Prof. Dr. Ahmed Rafea, External Examiner, Professor at American
University in Egypt**

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2021

Engineer's Name: Kariman Mahoud Hamouda
Date of Birth: 19/10/1992
Nationality: Egyptian
E-mail: Kariman_Elshakankery@hotmail.com
Phone: 01111246043
Address: Nasr City
Registration Date: 1/10/2014
Awarding Date: .../.../...2021.....
Degree: Master of Science
Department: Computer Engineering



Supervisors:
Prof. Dr. Magda Bahaa ElDin Fayek, Thesis Main Advisor
Dr. Mona Farouk Ahmed, Advisor

Examiners:
Prof. Dr. Magda Bahaa ElDin Fayek, Thesis Main Advisor
Prof. Dr. Nevin Mahmoud Darwish, Internal Examiner
Prof. Dr. Ahmed Rafea, External Examiner, Professor at American University in Egypt

Title of Thesis:

HILATSA: A Hybrid Incremental Learning Approach for Arabic Tweets Sentiment Analysis

Key Words:

Sentiment Analysis, Opinion Mining, Hybrid Approach, Arabic Twitter Sentiment Analysis, Sentiment Classification

Summary:

Sentiment analysis (SA) is the process of analyzing writers' opinions, emotions and attitudes from documents and determining whether they are positive, negative or neutral. It is used for many reasons as developing product quality, adjusting market strategy and improving customer services. Since the evolution in technology and the tremendous growth of social networks, a huge amount of data is generated. In spite of the availability of data, there is a lack of tools and resources. Though Arabic is a popular language, there are too few dialectal Arabic analysis tools. This is because of the many challenges in Arabic due to its morphological complexity and its dynamic nature. Approaches used to classify the opinions are categorized into lexicon based, machine based and hybrid based approach. This thesis introduces a semi-automatic learning system for sentiment analysis that is capable of updating the lexicon in order to be up to date with language changes. It is a hybrid approach which combines both lexicon based and machine learning approaches in order to identify Arabic tweets sentiments polarities. It proved to be able to cope with the dynamic nature and improved the accuracy by 17.55%.

Disclaimer

I hereby declare that this thesis is my own original work and that no part of it has been submitted for a degree qualification at any other university or institute.

I further declare that I have appropriately acknowledged all sources used and have cited them in the references section.

Name: Kariman Mahmoud Hamouda Elshakankery

Date: ../../...

Signature:

Acknowledgements

Thanks to Allah for everything. Thanks to my family for their patience, support and understanding. Thanks to my supervisors especially Dr Mona for her guidance and patience.

Table of Contents

ACKNOWLEDGEMENTS.....	I
TABLE OF CONTENTS.....	III
LIST OF TABLES	VI
LIST OF FIGURES	VII
LIST OF ACRONYMS.....	VIII
ABSTRACT	IX
CHAPTER 1 : INTRODUCTION	1
1.1. OVERVIEW	1
1.2. MOTIVATION.....	1
1.3. PROBLEM STATEMENT	2
1.4. PURPOSE	2
1.5. ORGANIZATION OF THE THESIS.....	2
CHAPTER 2 : BACKGROUND.....	3
2.1. INTRODUCTION	3
2.2. DATASETS	3
2.3. PREPROCESSING	4
2.3.1. Common Steps.....	4
2.3.1.1. Tokenization	4
2.3.1.2. Normalization	4
2.3.1.3. Removing Stop Words.....	4
2.3.1.4. Detecting Negation	4
2.3.1.5. Detecting Intensification.....	5
2.3.1.6. Stemming.....	5
2.3.1.7. Part of Speech Tagging (POS).....	5
2.3.2. Social Media Based Preprocessing.....	5
2.3.3. Arabic Based Preprocessing	5
2.4. FEATURE SELECTION\TRANSFORMATION.....	6
2.4.1. Pointwise Mutual Information (PMI)	6
2.4.2. Chi Square χ^2	6
2.4.3. Information Gain (IG)	6
2.4.4. Latent Semantic Indexing (LSI)	7
2.4.5. Feature Selection for Sentiment Analysis Based on Content And Syntax Models	7
2.4.6. Gini-Index (GI).....	7
2.5. CLASSIFIERS	8
2.5.1. Naïve Bayes (NB).....	8
2.5.2. Complement Naïve Bayes (CNB)	8
2.5.3. K-Nearest Neighbor (KNN)	8
2.5.4. Support Vector Machine (SVM)	9
2.5.5. Logistic Regression (LR).....	10
2.5.6. Decision Tree.....	10

2.5.7. Neural Network (NN)	10
2.5.8. Convolutional Neural Network (CNN)	11
2.5.9. Recurrent Neural Networks (RNN)	11
2.5.10. Long Short Term Memory (LSTM)	11
2.6. EVALUATION METRICS	12
CHAPTER 3 : RELATED WORK.....	13
3.1. INTRODUCTION	13
3.2. DATASETS	13
3.2.1. ASTD: Arabic Sentiment Tweets Dataset	13
3.2.2. Twitter Dataset for Arabic Sentiment Analysis.....	13
3.2.3. Emoticon Smoothed Language Model (ESLAM)	13
3.2.4. Arasenti-Tweet	13
3.2.5. Arabic Jordanian General Tweets (AJGT)	14
3.2.6. A Large Scale Arabic Book Reviews Dataset (LABR).....	14
3.2.7. ArSAS: An Arabic Speech-Act and Sentiment Corpus of Tweets.....	14
3.2.8. An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis (GS).....	14
3.2.9. Syrian Tweets Corpus.....	14
3.3. LEXICON BASED APPROACH	14
3.3.1. Arabic Sentiment Analysis: Lexicon-Based and Corpus-Based.....	15
3.3.2. Lexicon-Based Methods for Sentiment Analysis	16
3.3.3. Contextual Semantics for Sentiment Analysis of Twitter	16
3.3.4. Sentiment Analysis on Tweets about Diabetes: An Aspect Level Approach.....	18
3.3.5. Contextual Sentiment Analysis for Social Media Genres	18
3.4. MACHINE BASED APPROACH	19
3.4.1. Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews	19
3.4.2. Analyzing Sentiment Level of Social Media Data Based on SVM And Naïve Bayes Algorithms	20
3.4.3. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts	20
3.4.4. Arabic Tweets Sentimental Analysis Using Machine Learning.....	21
3.4.5. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements	21
3.4.6. Large Scale Sentiment Analysis on Twitter with Spark.....	21
3.5. HYBRID BASED APPROACH.....	22
3.5.1. Sentiment Analysis for Modern Standard Arabic and Colloquial	22
3.5.2. Sentiment Analysis in Arabic Tweets	23
3.5.3. Arabic Tweets Sentiment Analysis-A Hybrid Scheme	23
3.5.4. A Web-Based Tool for Arabic Sentiment Analysis	24
3.5.5. Twitter Sentiment Analysis for Arabic Tweets	24
3.5.6. Arabic Sentence Level Sentiment Analysis.....	25
3.5.7. Arabic Language Sentiment Analysis on Health Services	26
3.5.8. Arabic Sentiment Analysis with Optimal Combination of Features Selection and Machine Learning Approaches.....	26
3.5.9. Combining Lexical Features and a Supervised Learning Approach for Arabic Sentiment Analysis	27
3.5.10. Hybrid Deep Learning for Sentiment Polarity Determination of Arabic Microblogs	27

3.5.11. A Sentiment Analysis Approach Based on Arabic Social Media Platforms	28
3.5.12. NTNUSentEval at SemEval-2016 Task 4: Combining General Classifiers for Fast Twitter Sentiment Analysis	28
3.5.13. A System for Real-Time Twitter Sentiment Analysis of 2012 US Presidential Election Cycle.....	29
3.5.14. Combining Lexicon-Based and Learning-Based Methods for Twitter Sentiment Analysis	29
3.5.15. Aspect Extraction for Opinion Mining with A Deep Convolutional Neural Network	30
3.5.16. Tweet Sentiment Analysis with Classifier Ensembles	30
3.5.17. Sentiment Analysis of Colloquial Arabic Tweets	31
3.5.18. Integrated Real-Time Big Data Stream Sentiment Analysis Service	31
3.5.19. Evaluating Social Context in Arabic Opinion Mining	31
3.5.20. Opinion Mining and Analysis for Arabic Language	32
3.6. SUMMARY	33
CHAPTER 4 : PROPOSED APPROACH.....	34
4.1. DATASETS	34
4.2. BUILDING LEXICONS.....	41
4.2.1 Words Lexicon	41
4.2.2. Emotions Lexicon.....	41
4.2.3. Intensification Words	42
4.2.4. Special Intensification Words.....	42
4.2.5. Idioms Lexicons	42
4.3. PREPROCESSING	43
4.4. FEATURES	44
4.4.1. Words' Weights.....	44
4.4.2. Building the Feature Vector.....	47
4.5. CLASSIFICATION	48
4.6. LEXICON SEMI-AUTOMATIC UPDATE	48
CHAPTER 5: EXPERIMENTAL RESULTS	50
5.1. FEATURES SELECTION.....	50
5.2. CLASSIFICATION MODULE	51
5.3. LEXICON SEMI-AUTOMATIC UPDATE MODULE	57
CHAPTER 6: CONCLUSION AND FUTURE WORK.....	60
6.1. CONCLUSION.....	60
6.2. CONTRIBUTIONS	60
6.3. FUTURE WORK	60
REFERENCES	61

List of Tables

Table I.1: List of Acronyms	viii
Table 4.1: MASTD Dataset Statistics.....	36
Table 4.2: LASTD Dataset Statistics.....	38
Table 4.3: Datasets Distributions	39
Table 4.4: Examples from Words Lexicon.....	41
Table 4.5: Examples from the Emotions Lexicon	42
Table 4.6: Examples from the Idioms Lexicon	42
Table 4.7: Levenshtein Distance Algorithm Results	47
Table 5.1: Results of E1 to E6.....	51
Table 5.2: LASTD Unbalanced 3-Class Classification.....	51
Table 5.4: Unbalanced 3-Class Classification Problem	53
Table 5.5: Balanced 3-Class Results	54
Table 5.6: Unbalanced 2-Class Results	55
Table 5.7: Balanced 2-Class Results	56
Table 5.8: Lexicon Update Results for LASTD	57
Table 5.9: Lexicon Update Results	58
Table 5.10: A sample of new words that are added to the lexicon.....	59
Table 5.11: A sample of tweets that pass the learner phases with HILATSA's classification and actual classifications	59
Table 5.12: A sample of tweets with their HILATSA's classification suggestions and their actual classifications.....	59

List of Figures

Figure 2.1: KNN Example.....	8
Figure 2.2: High C value that leads to a small margin	9
Figure 2.3: Low C value that leads to a wide margin.....	10
Figure 2.4: Decision Tree Example.....	11
Figure 3.1: Feature Vector for 3.4.6	25
Figure 3.2: AOPI Lexicon-based Algorithm	32
Figure 3.3: Sentiment Intensity Algorithm.....	33
Figure 4.1: HILATSA's Block Diagram	35
Figure 4.3: MASTD Tweets N-Grams	37
Figure 4.4: LASTD Tweets Length.....	39
Figure 4.5: LASTD Tweets N-Grams	40
Figure 4.6: Negation Detection Algorithm.....	45
Figure 4.7: Detecting Negation Pattern Algorithm	46
Figure 4.8: Lexicon Semi-Automatic Update Algorithm.....	49

List of Acronyms

Table I.1: List of Acronyms

Acronym	Definition
API	Application Program Interface
CBOW	Continues Bag of Words
CNB	Complement Naïve Bayes
CNN	Convolutional Neural Network
DDO	Diabetes Diagnosis Ontology
DT	Decision Tree
FN	False Negative
FP	False Positive
KNN	K Nearest Neighbor
LR	Logistic Regression
LSTM	Long Short Term Memory
MSA	Modern Standard Arabic
NB	Naïve Bayes
NN	Neural Network
POS	Part of Speech Tagging
RF	Random Forest
RNN	Recurrent Neural Networks
SA	Sentiment Analysis
SG	Skip Gram
SVM	Support Vector Machine
SWN	SentiWordNet Lexicon
TDOC	Term Degree of Correlation
tf	Term Frequency
tf-idf	Term Frequency Inverse Document Frequency
TN	True Negative
TP	True Positive
URL	Universal Resource Locator