

**ENHANCED GRAY WOLF ALGORITHM BASED
FEATURE SELECTION FOR SENTIMENT
ANALYSIS**

TAMARA AMJAD ABDELKARIM ALQABLAN

UNIVERSITI SAINS MALAYSIA

2025

ENHANCED GRAY WOLF ALGORITHM BASED FEATURE SELECTION FOR SENTIMENT ANALYSIS

by

TAMARA AMJAD ABDELKARIM ALQABLAN

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

August 2025

ACKNOWLEDGEMENT

I am thankful to Allah SWT for giving me the strength to finish this study. My grateful thanks go to my supervisor Dr. Mohd Halim Bin Mohd Noor, whose wise counsel, unwavering support, encouragement, vast knowledge, and valuable insights have been indispensable throughout this journey. I am also sincerely thankful to my co-supervisor, Professor Dr. Mohammed Azmi Al-Betar, for his helpful advice, insightful comments, immense knowledge, motivation and steadfast support.

My family deserves special thanks. Words cannot express how grateful I am to my father, and mother, for all the sacrifices that they have made for me. Finally, I thank all my friends who encouraged me throughout the duration of this study.

TABLE OF CONTENTS

ACKNOWLEDGEMENT	ii
TABLE OF CONTENTS.....	iii
LIST OF TABLES	viii
LIST OF FIGURES	xiv
LIST OF ABBREVIATIONS	xvi
ABSTRAK.....	xvii
ABSTRACT	xix
CHAPTER 1 INTRODUCTION.....	1
1.1 Background	1
1.2 Research Motivation.....	5
1.3 Problem Statement	7
1.4 Research Questions	10
1.5 Research Objectives	10
1.6 Research Significance.....	11
1.7 Research Scope	11
1.8 Research Organization.....	12
CHAPTER 2 LITERATURE REVIEW	13
2.1 Introduction.....	13
2.2 Sentiment Analysis Background	13
2.3 Stages of Sentiment Analysis.....	15
2.3.1 Pre-Processing Stage.....	15
2.3.2 Feature Extraction Stage	16
2.3.2(a) Bag of Words.....	17

2.3.2(b)	Term Frequency-Inverse Document Frequency	17
2.3.2(c)	Word Embedding	18
2.3.3	Feature Selection Stage	19
2.3.4	Modeling Stage	20
2.3.4(a)	Lexicon-Based Model	20
2.3.4(b)	Machine Learning-Based Model	23
2.3.4(c)	Deep Learning-Based Model	25
2.4	Arabic Sentiment Analysis	28
2.5	Feature Selection	32
2.5.1	Filter-Based Method	33
2.5.2	Wrapper-Based Method	34
2.5.3	Hybrid-Based Method	35
2.6	Metaheuristics Techniques	37
2.7	Swarm Metaheuristics-based FS in SA	38
2.8	GWO-based FS Approaches	43
2.9	HHO-based FS Approaches	47
2.10	Critical Analysis	54
2.11	Conclusion	59
CHAPTER 3 METHODOLOGY		61
3.1	Introduction	61
3.2	Research Methodology Phases	61
3.3	Text Preprocessing Phase	63
3.3.1	Tokenizing	63
3.3.2	Cleaning of Data	63
3.3.3	Stemming	64

3.3.4	Analyzing Emoji	64
3.4	Text Preparation Phase	65
3.5	Construction phase	66
3.6	Improvement phase	68
3.6.1	BGWO Algorithm.....	69
3.6.2	BHHO Algorithm.....	72
3.6.3	The Adaptive β -hill Climbing Algorithm.....	77
3.6.4	Hybrid filter-wrapper sentiment feature selection approaches.....	79
3.6.4(a)	ReliefF-A β -BGWO Approach.....	79
3.6.4(b)	ReliefF-A β -BGWHHO Approach	80
3.7	Standard Datasets	81
3.7.1	UCI Reviews Datasets.....	82
3.7.2	Twitter Dataset.....	82
3.7.3	OCA Dataset.....	82
3.7.4	Arabic Twitter DataSet (Ar-Twitter)	83
3.7.5	Software Dataset	83
3.7.6	Political Dataset	83
3.8	Evaluation	83
3.8.1	Accuracy.....	84
3.8.2	Precision	84
3.8.3	Recall.....	84
3.8.4	F-measure	85
3.8.5	Friedman Test.....	85
3.8.6	Holm's Test	85
3.9	Conclusion	85

CHAPTER 4 IMPROVED BINARY GRAY WOLF OPTIMIZER BASED ON ADAPTIVE B-HILL CLIMBING FOR SENTIMENT FEATURE SELECTION.....	87
4.1 Introduction.....	87
4.2 Construction Phase	87
4.2.1 Objective Function.....	89
4.2.2 Initial Solution Generation and Representation.....	90
4.2.3 Improvement Phase: A β -BGWO	93
4.3 Evaluation Phase	94
4.3.1 Datasets	94
4.3.2 Experimental Setup.....	95
4.3.3 Results and Discussion.....	96
4.4 Conclusion	108
CHAPTER 5 GRAY WOLF HARRIS HAWK ALGORITHM FOR FEATURE SELECTION PROBLEM	111
5.1 Introduction.....	111
5.2 Construction Phase	111
5.2.1 Objective Function.....	112
5.2.2 Initial Solution Generation and Representation.....	113
5.2.3 Improvement Phase: BGWHHO	113
5.3 Evaluation Phase	116
5.3.1 Experimental Setup.....	116
5.3.2 Results and Discussion.....	117
5.4 Conclusion	128
CHAPTER 6 A NEW GRAY WOLF HARRIS HAWK OPTIMIZATION ALGORITHM FOR ENHANCING FEATURE SELECTION.130	
6.1 Introduction.....	130

6.2	Construction Phase	130
6.2.1	Objective Function.....	131
6.2.2	Initial Solution Generation and Representation.....	131
6.2.3	Improvement Phase: $A\beta$ -BGWHHO	131
6.3	Evaluation Phase	136
6.3.1	Experimental Setup	136
6.3.2	Results and Discussion.....	138
6.4	Conclusion	169
CHAPTER 7 CONCLUSION AND FUTURE WORK.....		171
7.1	Introduction.....	171
7.2	Research Summary.....	171
7.3	Contributions Against Objectives	172
7.4	Future Research.....	174

LIST OF TABLES

	Page
Table 2.1 Example of <i>BoW</i> Representation of Two Reviews.....	17
Table 2.2 Summary of the Advantages and Disadvantages of SA Models	26
Table 2.3 Hybrid Approaches	36
Table 2.4 Optimization Approaches Summary.....	42
Table 2.5 GWO-Based Feature Selection Summary.....	47
Table 2.6 HHO-based Feature Selection Summary	53
Table 2.7 Summary of critical analysis of related works	57
Table 3.1 Impact of Emoji Conversion	65
Table 3.2 Text Preprocessing and Text Preparation Steps	66
Table 3.3 Numerical example illustrating the construction phase steps.	68
Table 3.4 Description of the datasets used in the experiment.	81
Table 3.5 UCI Dataset	82
Table 4.1 Feature Selection Benchmark Datasets.....	95
Table 4.2 A β -BGWO Parameters.	96
Table 4.3 A Comparison Between the Proposed A β -BGWO and the Original Dataset Using the KNN Classifier in Terms of Accuracy and No. of Selected Features.	98
Table 4.4 Classification Accuracy Comparison Between the Proposed A β - BGWO, the Two Approaches of BGWO (bGWO1, bGWO2), and Original Dataset Using KNN Classifier.....	98
Table 4.5 Comparison of Classification Accuracy Obtained by A β -BGWO with State-Of-The-Art Feature Selection Approaches for 18 UCI	

	Datasets.	102
Table 4.6	Comparison of the Number of Selected Features Obtained by A β -BGWO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.....	102
Table 4.7	Comparison Results in Terms of the Fitness Values Obtained by A β -BGWO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.....	106
Table 4.8	Average rankings in terms of the number of features for all competitor approaches using Friedman's test.....	107
Table 4.9	Holm's results between A β -BGWO and another in terms of the number of selected features.....	107
Table 4.10	Average rankings in terms of accuracy for all competitor methods using Friedman's test.	108
Table 4.11	Holm's results between AB-BGWO and another in terms of accuracy.	108
Table 5.1	BGWHHO Parameters.....	117
Table 5.2	A Comparison Between the BGWHHO Approaches with the Original Datasets in Terms of Accuracy and No. of Selected Features.	118
Table 5.3	Classification Accuracy Comparison Between the BGWHHO Approach, the Two Approaches of BGWO (bGWO1, bGWO2), and the Original Dataset Using KNN Classifier.	118
Table 5.4	Exploration and Exploitation Behaviours of the BGWO, BHHO, and BGWHHO Algorithms Across UCI Datasets.	119
Table 5.5	Comparison of Classification Accuracy Obtained by the BGWHHO Approach with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.....	122

Table 5.6	Comparison of the Number of Selected Features Obtained by the BGWHHO Approach with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	123
Table 5.7	Comparison Results in Terms of the Average Fitness Values Obtained by BGWHHO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	124
Table 5.8	Average rankings in terms of the number of features for all competitor approaches using Friedman’s test.	127
Table 5.9	Holm’s results between β HC and another in terms of the number of selected features.	127
Table 5.10	Average rankings in terms of accuracy for all competitor methods using Friedman’s test.	128
Table 5.11	Holm’s results between $A\beta$ -BGWO and another in terms of accuracy.	128
Table 6.1	Test Benchmark Optimization Functions	137
Table 6.2	$A\beta$ -BGWHHO Parameters.	138
Table 6.3	Sentiment Analysis Benchmark Datasets	138
Table 6.4	A Comparison of Different Algorithms on Numerical Optimization Problems.	140
Table 6.5	Exploration and Exploitation of the Two Proposed Algorithms on Numerical Optimization Problems.	143
Table 6.6	A Comparison Between the Proposed $A\beta$ -BGWHHO, $A\beta$ -BGWO, and BGWHHO Approaches with the Original Datasets Using the KNN Classifier in Terms of Accuracy and No. of Selected Features.	144

Table 6.7	Classification Accuracy Comparison Between the A β -BGWHHO, A β -BGWO, BGWHHO, the Two Approaches of BGWO (bGWO1, bGWO2), and the Original Dataset Using KNN Classifier.	145
Table 6.8	Exploration and Exploitation of the Proposed Algorithms Across UCI datasets.	146
Table 6.9	Comparison of Classification Accuracy Obtained by the A β -BGWHHO Approach with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	153
Table 6.10	Comparison of the Number of Selected Features Obtained by the A β -BGWHHO Approach with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	154
Table 6.11	Comparison Results in Terms of the Average Fitness Values Obtained by A β -BGWHHO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	155
Table 6.12	Experimental Outcomes of ReliefF-A β -BGWHHO in Terms of Accuracy, Number of Selected Features, Fitness, Precision, Recall, and F-measure for All Datasets.	156
Table 6.13	Comparison of Classification Accuracy Obtained by A β -BGWHHO with State-Of-The-Art Algorithms for Arabic Datasets.	157
Table 6.14	Comparison of the Number of Selected Features Obtained by A β -BGWHHO with State-Of-The-Art Algorithms for the Arabic dataset.....	157
Table 6.15	Comparison Results in Terms of the Average Fitness Values Obtained by A β -BGWHHO with State-Of-The-Art Algorithms for the Arabic dataset.....	158

Table 6.16	Comparison of Classification Accuracy Obtained by $A\beta$ -BGWO and $A\beta$ -BGWHHO with State-Of-The-Art Algorithms for English Datasets.	158
Table 6.17	Comparison of the Number of Selected Features Obtained by $A\beta$ -BGWHHO with Some State-Of-The-Art Algorithms for English Datasets.	159
Table 6.18	Comparison Results in Terms of the Average Fitness Values Obtained by $A\beta$ -BGWHHO with State-Of-The-Art Algorithms for English Datasets.	159
Table 6.19	Average rankings in terms of the number of features for all competitor approaches using Friedman's test for 18 UCI Datasets.	160
Table 6.20	Holm's results between $A\beta$ -BGWHHO and another in terms of the number of selected features for 18 UCI Datasets.	161
Table 6.21	Average rankings in terms of accuracy for all competitor methods using Friedman's test for 18 UCI Datasets	161
Table 6.22	Holm's results between $A\beta$ -BGWHHO and another in terms of accuracy for 18 UCI Datasets.....	162
Table 6.23	Average rankings in terms of number of selected features for all competitor approaches using Friedman's test for Arabic and English datasets	163
Table 6.24	Holm's results between $A\beta$ -BGWHHO and another in terms of the number of selected features for Arabic and English datasets.	163
Table 6.25	Average rankings in terms of accuracy for all competitor approaches using Friedman's test for Arabic and English datasets.....	164
Table 6.26	Holm's results between $A\beta$ -BGWHHO and another in terms of accuracy for Arabic and English datasets.	164
Table 6.27	Comparison Between the Ensemble Approach, KNN, SVM, NB, RF, and DT using Features Selected by $A\beta$ -BGWHHO in Terms of	

	Accuracy, Precision, Recall, F-measure, and Number of Selected Features	167
Table 6.28	Comparative Results.	168

LIST OF FIGURES

	Page
Figure. 1.1 Integration of Sentiment Analysis IoT Device.....	5
Figure. 2.1 Flowchart Representation of Pre-Processing Operations.....	16
Figure 2.2 Sentiment Analysis Models.....	20
Figure 2.3 Flowchart for Lexicon-Based Model.....	21
Figure 2.4 Flowchart for Machine Learning-Based Model	24
Figure 2.5 Flowchart for Deep Learning-Based Model.....	26
Figure 3.1 Overview of Research Methodology.	62
Figure 4.1 A β -BGWO Flowchart.	88
Figure 4.2 Solution Representation	91
Figure 4.3 Initial Solution Representation	93
Figure 4.4 The Convergence Curves of the A β -BGWO Algorithm with State-Of-The-Art Algorithms for All Datasets.	99
Figure 4.5 Classification Accuracy Obtained by A β -BGWO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	101
Figure 4.6 Number of Selected Features Obtained by A β -BGWO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	105
Figure 5.1 Average Number of Selected Features Obtained by BGWHHO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	123
Figure 6.1 Average Number of Selected Features Obtained by A β -BGWHHO with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets.	150

Figure 6.2 The Convergence Curves of the $A\beta$ -BGWHHO Approach with State-Of-The-Art Feature Selection Approaches for 18 UCI Datasets..... 151

LIST OF ABBREVIATIONS

A β HC	Adaptive β Hill Climbing
BGWO	Binary Gray Wolf Optimization
BHHO	Binary Harris Hawk Optimization
Bi-LSTM	Bidirectional LSTM
CBOW	Continuous Bag-Of-Words
CNN	Convolution Neural Networks
DF	Document Frequency
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
FE	Feature Extraction
FS	Feature Selection
IDC	International Data Corporation
IDF	Inverse Document Frequency
IG	Information Gain
IoT	Internet of Things
KNN	K-Nearest Neighbors
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
NP	Nondeterministic Polynomial
RF	Random Forst
RNN	Recurrent Neural Networks
SA	Sentiment Analysis
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency

PEMILIHAN CIRI BERASASKAN ALGORITMA SERIGALA KELABU DIPERTINGKAT UNTUK ANALISIS SENTIMEN

ABSTRAK

Orang ramai mengakses platform media sosial melalui pelbagai peranti IoT, termasuk telefon pintar, tablet dan peranti boleh pakai, yang membawa kepada penjana kuantiti data yang ketara. Data berdimensi tinggi dianggap sebagai cabaran utama dalam menganalisis dan mengekstrak pengetahuan daripada data raya. Pemilihan ciri adalah teknik yang berkaitan dengan memilih hanya ciri-ciri yang paling relevan untuk meningkatkan prestasi. Dalam konteks pendekatan pemilihan ciri berasaskan pembalut, algoritma sering terperangkap dalam penyelesaian yang lemah (subset ciri) kerana ketidakupayaan untuk meneroka ruang carian tempatan dengan cekap. Untuk menangani masalah ini, pertama, pendekatan berasaskan pembalut yang ditingkatkan yang menggabungkan algoritma BGWO dengan algoritma carian tempatan A β HC (A β -BGWO) telah dicadangkan untuk meningkatkan keupayaan penerokaan BGWA dan meningkatkan prestasinya dalam hal keupayaan carian setempat. Kedua, mengukuhkan A β -BGWO dengan mengintegrasikannya dengan BHHO (A β -BGWHHO) untuk meningkatkan penumpuan global A β -BGWO, meningkatkan keupayaan meneroka, dan menyesuaikan antara pengeksploitasi dan penerokaan. Ketiga, kerana sifat set data analisis sentimen, kaedah penapis boleh secara berkesan mengenal pasti dan menghapuskan ciri-ciri hingar tersebut, dengan itu meningkatkan kualiti subset ciri yang dinilai oleh pendekatan pembalut. Untuk menangani masalah ini, dua pendekatan penapis-pembalut

telah dicadangkan, menggabungkan kaedah penapis ReliefF dengan kedua-dua pendekatan berasaskan pembalut yang ditingkatkan. Untuk proses penilaian, pendekatan pembalut yang dicadangkan dinilai dan dibandingkan dengan pendekatan terkini menggunakan 18 set data UCI. Hasilnya menunjukkan bahawa pendekatan yang dicadangkan secara ketara meningkatkan eksploitasi dan penyelidikan BGWO dan melebihi pelbagai pendekatan manakala pendekatan penapis-pembalut yang disyorkan dinilai menggunakan lapan set data bahasa Arab dan Inggeris yang berbeza dan dibandingkan dengan satu set pendekatan dalam literatur. Untuk meningkatkan hasil klasifikasi, model gabungan telah dibangunkan menggabungkan K-Nearest Neighbors (KNN), Support Vector Machines (SVM), dan Naive Bayes. (NB). Hasil menunjukkan bahawa pendekatan ReliefF-wrapper yang dicadangkan memperoleh hasil terbaik pada 7 set data dalam ketepatan, ketepatan, pengingat, dan F-pengukuran.

ENHANCED GRAY WOLF ALGORITHM BASED FEATURE SELECTION FOR SENTIMENT ANALYSIS

ABSTRACT

People access social media platforms through a range of IoT devices, including smartphones, tablets, and wearable devices, which lead to the generation of a significant quantity of data. The high dimensional of data is considered a major challenge in analyzing and extracting knowledge from big data. Feature selection is a technique concerned with selecting only the most relevant features to improve performance. In the context of wrapper-based feature selection approaches, the algorithm is often stuck in a poor solution (feature subset) due to its inability to efficiently explore the search space. To tackle this problem, first, an enhanced wrapper-based approach that combines the BGWO algorithm with the A β HC local search algorithm (A β -BGWO) was proposed to enhance the BGWO exploitation capabilities and improve its performance in terms of local search ability. Second, enhance the A β -BGWO by integrating it with BHHO (A β -BGWHHO) to improve the global convergence of A β -BGWO, enhance its exploration capabilities, and fine-tune between exploitation and exploration. Third, due to the nature of the sentiment analysis datasets, filter methods can effectively identify and eliminate noise features, thereby enhancing the quality of feature subsets evaluated by the wrapper approach. To tackle this issue two filter-wrapper approaches have been proposed, combining the ReliefF filter method with the two enhanced wrapper-based approaches. For the evaluation process, the proposed wrapper approaches are evaluated and compared with the state-of-the-art approaches using 18 UCI datasets. Results showed that the proposed approaches

notably enhanced the exploitation and exploration of the BGWO and outperformed a wide range of approaches whereas the proposed Filter-wrapper approaches were evaluated using eight different Arabic and English datasets and compared with a set of approaches in the literature and with three different filter methods (CHI, Gain Ratio, and IG) when integrated with the proposed wrapper approaches. To improve classification results, an ensemble model has been developed combining K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes (NB). Results show that the proposed ReliefF-wrapper approaches obtained the best results on 7 datasets in accuracy, precision, recall, and F-measure.

CHAPTER 1

INTRODUCTION

1.1 Background

We live in an era defined by digital connectivity between several devices via the Internet, the Internet of Things (IoT) is the convergence of electronic appliances and devices with Internet access that can exchange data and provide useful insights (Agrawal & Moparthi, 2023; Zhou et al., 2018). The growing use of the Internet and the rise of social media platforms like Twitter, Facebook, Instagram, etc. have offered an ideal environment for people to express their opinions in a variety of areas, including politics, economics, arts, etc. These platforms are used by people to keep up with the latest news and services and to share their comments and recommendations of products, services, and policies that they use or are interested in in their lives. People access these social media platforms through a range of IoT devices, including smartphones, tablets, and wearable devices, which lead to the generation of a significant quantity of data.

The Sentiment Analysis (SA) community is a subfield of the Natural Language Processing (NLP) research communities, has focused on analyzing people's thoughts on a given issue on these platforms to take a wider overview, which may aid decision-makers to take suitable strategies and decisions based on these opinions.

Many NLP experts are interested in this area and have begun investigating the uses of SA. Various research communities are now striving to further this field, from straightforward linear approaches to Machine Learning (ML) approaches. However, a lot of studies have concentrated on English because it is the dominant language of science and other languages have fallen behind.

An increase in Arabic-language digital material, notably on social media, has coincided with a sharp rise in the number of internet users (In March 2017, there were about 141 million Arabic internet users.). Due to the extensive use of Arabic in social networks, this has boosted the demand for Arabic SA.

SA problem is defined as a "suitcase" problem that requires the completion of several NLP sub-tasks such as text preprocessing, feature extraction, feature selection and sentiment classification (Fang & Zhan, 2015; Keerthi Kumar et al., 2019). There are several studies of SA have been conducted on a variety of input data formats including text SA, audio SA and video SA (Chaturvedi et al., 2019; Poria et al., 2017). In general, text SA was used in the majority of these studies. Popularly, Text SA aims to classify texts either into two categories: positive or negative (Alharbi & de Doncker, 2019), or into three categories: positive, negative, and neutral (Junior et al., 2021). Furthermore, other studies classified texts into several categories based on the study's topic (Bouazizi & Ohtsuki, 2019).

Most SA approaches can be categorized into lexicon-based approaches, ML-based approaches, and hybrid approaches (Birjali et al., 2021; Patel & Thakkar, 2015; Sankar H & Subramaniaswamy V, 2017). Traditional (lexicon-based) approaches have been used, which are a list of lexical features labeled according to their sentiment polarity positive or negative (S. Kumar et al., 2019), these approaches are separated into two types: Dictionary-based and Corpus-based (Gupta & Agrawal, 2020). After that, the ML-based approaches came to overcome the shortcomings of lexicon-based approaches. ML-based approaches can be divided into three categories: supervised, unsupervised, and semi-supervised approaches (Chandra Pandey et al., 2017), whereas the hybrid approaches

merge more than one approach, for instance, a combination of lexicon-based and ML-based approaches, or a combination of lexicon-based and deep learning-based approaches.

The field of optimization is significant and has numerous essential applications, where the algorithms used for optimization are varied and have a variety of successful uses. Metaheuristic algorithms are general approaches (templates) that can be applied as a model in designing fundamental heuristics to solve specific optimization problems. A broad variety of metaheuristic algorithms have appeared and most of them are nature-inspired (Akinola et al., 2022; Talbi, 2009). Modern metaheuristics are becoming more and more common among optimization approaches, giving rise to a new subfield of optimization termed metaheuristic optimization (Hussain et al., 2019; Yang, 2012), metaheuristic optimization algorithms have been developed with an aim to carry out global search with three main purposes: solving problems faster, solving large problems, and obtaining robust algorithms. The efficiency of metaheuristic algorithms can be attributed to the fact that they mimic the best characteristics of nature, especially the selection of the fittest in biological systems which have evolved by natural selection over millions of years (Ezugwu et al., 2021; Gandomi et al., 2013; Oyelade et al., 2022). These metaheuristic algorithms were developed and applied in almost all optimization domains, like design, scheduling and planning, data mining, machine intelligence, and many others.

Metaheuristic optimization approaches have also been employed for SA. Some researchers employed metaheuristic approaches to reduce the number of features in order to enhance SA performance (Hosseinalipour & Ghanbarzadeh, 2022; Iqbal et al., 2019), while others used it to find the optimal clusters (Chandra Pandey et al., 2017; Shekhawat et al., 2021).

Most of the social media reviews have irrelevant, noisy, and redundant data. So, the SA model needs to select the subset of sentiment features from these data. Sentiment data selection is concerned with minimizing data size and picking only the most relevant features in order to improve SA results and lower computing costs. Based on their selection mechanisms, sentiment feature selection approaches can be categorized into Filter-based feature selection, Wrapper-based feature selection, and Hybrid feature selection (Guyon & De, 2003; Hosseinalipour & Ghanbarzadeh, 2022; Madasu & Elango, 2020; N. Hoque et al., 2016).

In the filter-based feature selection approach, features are evaluated using statistical metrics, ignoring the inter-dependencies between them. The highest-ranking features are selected, which raises the likelihood of discovering irrelevant and redundant features that lead to decreased performance. ReliefF, Chi-square (CHI), Information Gain (IG), and Gain ratio are examples of common filter approaches (Hassonah et al., 2020; Labani et al., 2018). The wrapper-based feature selection approach is a combination of learning algorithms and feature subset generation strategies, it evaluates a candidate feature subset of features dependent on the effectiveness of the applied learning algorithm (Alweshah et al., 2022; H. Liu et al., 2019), where the feature subset is formed according to the opted search strategy, for example, forward selection or backward selection (Urbanowicz et al., 2018).

Generally, hybrid methods combine diverse approaches to give results that are more accurate and obtain the best feasible sentiment feature subset. This method tries to combine the strengths of models (e.g., wrapper and filters) by utilizing their different evaluation criteria at various phases of the search process (Ansari et al., 2019; H. Liu &

Yu, 2005). Usually, two or more feature selection methods of various conceptual origins are combined (Remeseiro & Bolon-Canedo, 2019). For instance, removing a set of features with a less computationally expensive filter method and then finding the best candidate subset of features with a more computationally expensive wrapper method.

Different metaheuristic strategies to solve the sentiment feature selection problem were implemented in order to get the best possible sentiment feature subset, these algorithms formulated the sentiment feature selection issue as an NP-hard optimization, so the use of metaheuristic strategies is very appropriate to solve the sentiment feature selection problem (Rohaidah Ahmad et al., 2019; Yusta, 2009; H. Zhao et al., 2021). Genetic Algorithm, Cuckoo Search and Particle Swarm Optimization are examples of these algorithms.

1.2 Research Motivation

Social network sentiment analysis can be integrated with social IoT devices to enhance the smart home environment making it more comfortable. Imagine that your smart home system includes IoT devices like digital music players, light bulbs, and thermostats. These gadgets include built-in sentiment analysis algorithms and are linked to your social media profiles as illustrated in Figure 1.1.

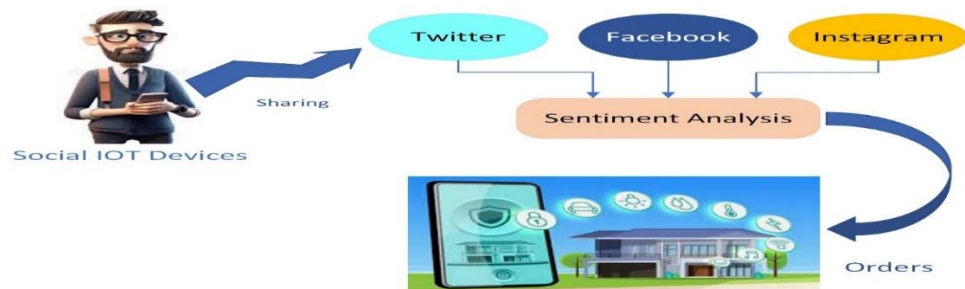


Figure. 1.1 Integration of Sentiment Analysis IoT Devices

The music player may use sentiment analysis to create personalized playlists based on your current state of mind. When someone shares something nostalgic on social media, a playlist of old favorites may be activated; when someone writes something enthusiastic, upbeat and lively music might be played. Smart lights can be set up throughout your house, and they will change color or brightness depending on the mood of your latest social media posts. For example, happy celebrations or positive social media posts may result in warm, welcoming hues such as orange or yellow, while a calming dimming effect could be produced by indicators of stress or sadness.

Furthermore, the smart thermostat can interpret user mood from temperature settings by studying social media activity. If your posts show a lot of motivation or enjoyment, the thermostat may be drooped a little bit to keep you feeling renewed. Conversely, if people start writing about how tired they are, it will raise the temperature and make you feel more comfortable.

This integration shows how sentiment analysis can be applied to social network data in IoT applications; moreover, it enables your smart home to adapt to your mood, making it more comfortable.

Another incentive regarding health and psychological monitoring. A variety of wearable health monitoring devices, including fitness trackers, smartwatches, and others, can be enhanced with the ability to detect mood. For example, if a user indicates feelings of stress or anxiety, the device could suggest relaxation techniques or breathing exercises. On the flip side, postings about being happy or satisfied might prompt motivational messages or reminders to maintain healthy habits. Along with helping people manage

stress and enhance their mental health, this integration showcases the potential of combining sentiment analysis of social media posts with Internet of Things devices to promote overall well-being in smart cities.

This section presents how technology can adapt and respond to human emotions in real time by integrating sentiment analysis with IoT devices to improve people's quality of life by creating more comfortable environments.

In the fast-evolving domains of machine learning and data mining, feature selection plays a vital role in determining the effectiveness of predictive models. Specifically, within sentiment analysis, datasets often have an extremely large number of features derived from textual data. This high dimensionality can lead to issues like higher computational expenses, overfitting, and decreased interpretability of models. Consequently, effective feature selection methods are essential for improving the performance of sentiment classification models.

1.3 Problem Statement

SA aims to reveal sentiments, emotions, and viewpoints in textual data. In domains such as customer feedback analysis, social media monitoring, and market research, its importance has increased dramatically. SA data is considered a high-dimensionality dataset, which typically consists of thousands of attributes (A. Kumar et al., 2022; Tubishat et al., 2019). However, most of the data which often include words, phrases, and syntactic or semantic qualities is irrelevant, noisy, and redundant which leads to reduced model performance and raised computing costs.

Feature selection is a common data reduction technique, which is used to tackle the "curse of dimensionality" in data analysis (Crespo Márquez, 2022). It is a vital step in

the preprocessing phase of ML and data mining that aims to select relevant and informative features from large datasets (Parashar et al., 2023). Accordingly, reduces space dimensionality and boosts ML performance in terms of accuracy, simplicity, speed, and data interpretation. Attaining optimum feature selection in SA presents a formidable challenge, especially when managing complex, large-scale datasets. Existing techniques frequently encounter difficulties in balancing exploration and exploitation in the search space, hence constraining their efficacy in effectively identifying the optimal subset of features. Traditional feature selection approaches (filtering approaches) often struggle to find the optimal subset of features due to the high dimensionality and complexity of textual data (Verma & Sahu, 2024). Therefore, an efficient feature selection strategy is required to handle the large and complex feature spaces of SA tasks.

Wrapper-based approaches are widely used to solve feature selection problems since they provide the capability to assess feature subsets by considering the performance of a certain ML algorithm, where the efficacy of the search process relies on the optimization algorithm's performance used to effectively explore and exploit the search space (Das et al., 2024). Exploitation refers to the process of searching around existing good solutions to further enhance them, whereas exploration involves looking in different and diverse parts of the solution space to find potentially better solutions (Trojovský & Dehghani, 2023). However, a crucial challenge in optimization algorithms is achieving a suitable balance between exploration and exploitation. Reaching the proper balance between these two factors is essential to prevent early convergence to local optimal solutions and to ensure an effective exploration of the solution space.

Despite improvement in wrapper-based feature selection approaches, most of the used metaheuristic algorithms suffer from stagnating in local optimum solutions and lack an effective balance between exploration and exploitation strategies, resulting in failure to sufficiently explore the feature space or efficiently leverage known important features, which lead to an inferior performance across many applications, such as SA.

Meanwhile, regardless of the effective use of BGWO in tackling optimization problems across several areas, stagnation in the local optima area still be a problem as a result of its swift exploitation, particularly when tackling problems that include a substantial number of features and several local solutions (Geng et al., 2025; Makhadmeh et al., 2022). The leadership hierarchy of the BGWO algorithm prioritizes exploitation over exploration by updating each solution according to the top three solutions in the population (α , β , and δ), this can lead to a scarcity of variation within the population, which restricts the algorithm to a limited portion of the search space and raises the probability of being stuck in local optima (Geng et al., 2025). As the algorithm lacks flexibility in effectively exploiting local search regions, an adaptive strategy is crucial for dynamically adjusting the search space. This prevents wolves from prematurely converging on local optima, improves the positions of the wolves, and enhances overall optimization performance. The adaptive local search algorithm enhances optimization accuracy in later stages by modifying the search process, which reduces GWO's sensitivity to local optima (Chen et al., 2023; Hu et al., 2024). Furthermore, the GWO convergence factor decreases progressively with each iteration, leading to a reduced focus on exploration as the algorithm progresses. If this factor declines too rapidly, the algorithm emphasizes exploitation above exploration (Faris et al., 2018; Makhadmeh et al., 2023).

1.4 Research Questions

The section aims to present various research questions related to improving a wrapper sentiment feature selection approach performance.

These questions include:

- How to enhance the local exploitation of the BGWO algorithm in optimizing solution quality?
- How can the search region of the BGWO algorithm be enhanced to effectively navigate diverse search spaces?
- How will the balance between exploitation and exploration strategies affect the development of the sentiment feature selection wrapper approach?

1.5 Research Objectives

The main objective is to develop an algorithm to refine exploitation and exploration mechanisms and fine-tune between these two mechanisms, aiming to develop an effective sentiment feature selection approach that enhances the SA classification approach. For that, this research has the following objectives:

- To hybridize the BGWO population-based algorithm with the adaptive β -hill climbing local search algorithm to enhance the local exploitation search capability of the BGWO algorithm in the sentiment feature selection space and ensuring that the algorithm converges more effectively towards optimal subsets of features
- To enhance the exploration searchability of BGWO through the selective implementation of BHHO, leveraging BHHO's effective exploration strategy to

improve the performance of the wrapper approach in navigating the sentiment feature selection space.

- To propose an adaptive BGWHHO-based feature selection approach by integrating adaptive proposed algorithm with efficient components from the BHHO algorithm to balance between exploitation and exploration search strategies to navigate the sentiment search space effectively

1.6 Research Significance

After the research objectives are achieved, this study will have the following main contributions:

- Hybridization of the BGWO with the $A\beta$ HC local search algorithm ($A\beta$ -BGWO) to enhance BGWO local search capability.
- Integrating the BGWO and BHHO (BGWHHO) to enhance the exploration search strategy of the BGWO algorithm.
- Hybridization of the proposed adaptive algorithm $A\beta$ -BGWO with the BHHO algorithm to balance between exploitation and exploration in the feature selection search space.

1.7 Research Scope

This study covers the SA preprocessing steps (i.e., sentiment data selection and dimension reduction techniques) for English and Arabic datasets, the metaheuristic algorithms (i.e., different versions of the proposed BGWO), and the ML approaches to deal with the SA classification problem and compares the performance of the proposed method against publicly available methods. The methods proposed in this study will apply

to a set of available English and Arabic SA datasets. These datasets with different sizes (i.e., number of reviews, number of attributes, and number of clusters) are characterized by high-dimensional informative and uninformative text features. Different measurements will be used to evaluate the proposed methods.

1.8 Research Organization

The rest of this study is organized as follows:

Chapter 2 presents a detailed and comprehensive background of SA, SA approaches, Arabic SA approaches, the related work in sentiment feature selection using metaheuristic approaches, and an overview of the recent studies of feature selection models.

Chapter 3 introduces the research methodology, which is adopted in this research. The chapter consists of five phases: the data preprocessing phase, the data preparation, the construction phase, the improvement phase, and the evaluation phase.

Chapters 4, 5, and 6 present the A β -BGWO, BGWHHO, and A β -BGWHHO wrapper approaches. The chapters provide a thorough explanation of the suggested methodology and the order in which the operations were carried out. Additionally, the chapters give comparisons between each method and the state-of-the-art methods and discuss the experiments and outcomes of all the suggested methods. Furthermore, Chapter 6 presents a comprehensive experimental comparison that examines the sentiment feature selection results of the proposed filter-wrapper approaches. In Chapter 7, the conclusion of the study findings is presented, along with other possible works for the future.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides a structured and comprehensive review of the literature relevant to sentiment analysis (SA), with a focus on feature selection (FS) and optimization techniques to enhance classification performance. The review begins by outlining the fundamental concepts of SA, followed by a discussion of its core processing stages, including data preprocessing, feature extraction, and feature selection. The chapter then delves into various FS methods, highlighting their role in improving computational efficiency and model performance. In pursuit of more effective FS, the literature on metaheuristic optimization algorithms is reviewed, with a specific focus on swarm intelligence-based techniques applied in SA contexts. The final sections delve into advanced metaheuristic approaches, particularly those based on the Gray Wolf Optimizer (GWO) and Harris Hawk Optimizer (HHO), which have demonstrated promising results in optimizing FS for classification tasks. Through this review, existing gaps and limitations are highlighted, establishing the need for an improved optimization-based FS approach.

2.2 Sentiment Analysis Background

This section presents a definition of SA and gives an overview of the basic analysis stages, including the pre-processing, Feature Extraction (FE), FS, and modeling stages. SA (also known as opinion mining) is an active research area in NLP (Messaoudi et al., 2022; Tang et al., 2015). It aims at identifying, extracting, and organizing sentiments from

user-generated texts, audio, or visual data in social networks. The importance of SA comes from its ability to give an overview or a general sentiment on a specific topic, potentially helping decision makers to make the right decisions. Many studies have presented various approaches to SA, each dealing with the sentiment classification problem differently. In summary, SA can be done at three levels: document, sentence, and aspect levels (L. Zhang et al., 2018). A brief description of each level is given below.

The task of the document level is to classify the whole document and extract an overall view. The following document is considered as an example, “Overall, I believe the most recent phone I purchased is excellent. It comes with a high-resolution camera. The phone is extremely slim and fits easily into a pocket. The price is a little high given the configurations. However, the phone seems really good” . The document as a whole appears to be more positive than negative. So, if one word must be used to describe the review, it is positive.

At the sentence level, the task is to examine the sentences and determine whether each one reflects a positive, negative, or neutral viewpoint. Considering the example presented in the above subsection, the document can be divided into five sentences, each of which can be classified as positive, negative, or neutral. Here, opinions become slightly more specific.

The aspect level aims to find sentiments concerning the particular aspects of words (Ma et al., 2019). For example, the following sentence is considered, “I like the new design of the Instagram App, but I hate its customer policy.” , the review is on design and customers policy, which are two aspects of the product Instagram App, and the review is positive for the design aspect and negative for customers policy. The task at this level aids

in determining exactly what people like or dislike. It concentrates on the characteristics of features rather than the sentiment of paragraphs. Implicit or explicit aspect extraction can be considered the core task for SA (Tubishat et al., 2018).

2.3 Stages of Sentiment Analysis

The generic SA process is illustrated in Figure 1.2. The basic stages involve pre-processing, FE, and FS. Then, the classification stage can be performed using different SA models (e.g., ML, Deep Learning (DL)). The stages present in this section are generally used for text analysis. However, with a few modifications, the same curriculum can be used for audio and image data. Knowing that DL approaches can skip the first three stages and go straight to classification, they can self-learn features from the dataset. By contrast, these approaches are complicated and computationally expensive. A brief illustration of the pre-processing stage, FE stage, FS stage, and modeling stage are presented below followed by Table 2.2 which presents a summary of the SA approaches and their advantages and disadvantages.

2.3.1 Pre-Processing Stage

The next stage after collecting data is the pre-processing stage. The importance of this stage comes from removing unnecessary data to lower the size of the input raw dimension, enhancing analysis. The data are cleaned by deleting unnecessary words, URLs, stop words, punctuations, and other elements that are required to be reduced before the feature extraction stage. Sets of pre-processing techniques were illustrated by (Angiani et al., 2016; Duong & Nguyen-Thi, 2021; Pradha et al., 2019). Figure 2.1 represents a flowchart of the common pre-processing operations:

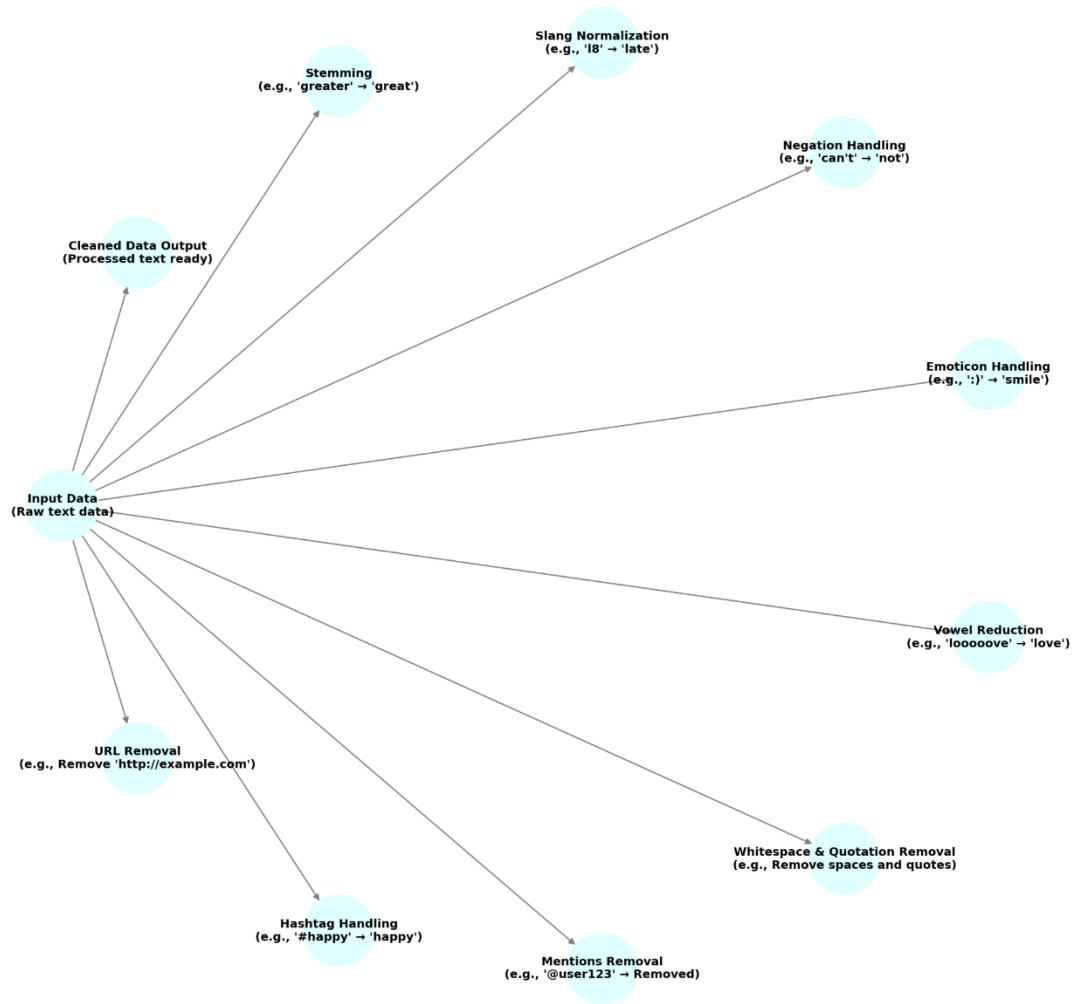


Figure. 2.1 Flowchart Representation of Pre-Processing Operations

2.3.2 Feature Extraction Stage

ML and DL models cannot process raw text data in its original form, therefore textual data needs to be converted into numerical representation suitable for these models. Text feature extraction can be defined as the process of extracting a list of valuable words from text data and converting them into a feature set that a classifier can use. The approaches listed below are commonly used to extract and represent features from text data.

2.3.2(a) Bag of Words

Bag of Words (BoW) is the most used technique for converting text to numerical representation, where a vector of integer numbers is generated representing the occurrence of words within a text. It is a simple way to represent the vector of data. However, the order or structure of words in a text is ignored while using this technique. *BoW* can be formulated as follows: let A and B two sentences $n(A)$ represent the set of words in sentence A , and $n(B)$ represent the set of words in sentence B then $BoW(A \cup B) = n(A) + n(B) - n(A \cap B)$. Suppose a new sentence C is added then $BoW(A \cup B \cup C) = n(A) + n(B) + n(C) - n(A \cap B \cap C)$.

For example, consider the following reviews: Review 1:” This car is very slow”, Review 2:” I like this car”. From Table 2.1, it can be observed that the vector length for all reviews is constant.

Table 2.1 Example of *BoW* Representation of Two Reviews

Reviews	This	Car	Is	Very	Slow	I	Like
Review 1	1	1	1	1	1	0	0
Review 2	1	1	0	0	0	1	1

2.3.2(b) Term Frequency-Inverse Document Frequency

Given that the BoW technique assesses the repetition of a word in a document without considering its relevance to the whole corpus, a modification on it called Term Frequency-Inverse Document Frequency (TF-IDF) was developed to calculate the word’s importance for the corpus.

The term frequency concept for word (i) (TF_{wi}) means the number of times a word (i) appears in a document d divided by the total number of words in the document. The

Inverse Document Frequency (IDF) concept is used to determine the significance of a word to a document in a set of documents (Ahuja et al., 2019; Waykole & Thakare, 2018).

For example: assuming a document” I like that car, which car do you mean? The white car” has 12 words and the word car appears three times in those 12 words, the term frequency will be $\frac{3}{12} = 0.25$. If there are 50,000 documents, only 500 of them contain the word “car”, then the $IDF_{car} = \log \frac{50000}{500} = 2$, and the $TF - IDF_{car} = 0.025 * 2 = 0.05$.

2.3.2(c) Word Embedding

Another common feature extraction technique is word embedding, which is a modeling language and feature learning method that converts words in a document into continuous real-number vectors. Word embedding is a distributed representation system of the documents’ vocabulary in which words with comparable meanings are similarly represented (L. Zhang et al., 2018b). Word2Vec, which was developed by (Mikolov et al., 2013), is a popular technique to learn word embedding. It includes two models, namely, skip-gram and Continuous Bag-Of-Words (CBOW). Both techniques depend on the likelihood of words occurring close together. Many other embedding methods like GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) were also used. A comparison between Word2Vec, Glove, Glovers, and FastText techniques was done in (H. N. Nguyen et al., 2021). The results showed that FastText outperformed other methods in terms of time consumption for training the classifier. Furthermore, a comparison between three DL models: Deep Neural Network (DNN), Convolution Neural Networks (CNN), and Recurrent Neural Networks (RNN) with TF-IDF and Word embedding for

different datasets revealed that combining DL techniques with word embedding is better than TF-IDF when performing SA (Dang et al., 2020).

The BoW technique is commonly employed due to its simplicity; however, they are often hindered by issues of high dimensionality, sparsity, and limited semantic representation. Although word embeddings provide richer contextual information, they require extensive computational resources and domain-specific adjustments (Kadhem et al., 2024). In contrast, TF-IDF offers a robust and efficient approach to text representation and preprocessing, addressing these challenges while maintaining computational feasibility.

2.3.3 Feature Selection Stage

A set of irrelevant and redundant features usually occurs among the set of features obtained in the previous stage. Therefore, FS techniques are needed to remove or reduce these undesirable features to reduce the computational cost and improve the performance of SA classification (T. H. Nguyen et al., 2015; Ramasamy & Meena Kowshalya, 2022). Many existing techniques were used like filter methods and wrapper methods. these methods will be reviewed in detail in section 2.5. The following lines give a brief description of the three main methods for selecting the features.

The filter method is designed in a way that makes feature selection and model learning independent (Birjali et al., 2017; Solorio-Fernández et al., 2020). features are evaluated using statistical metrics ignoring the fact that the inter-dependencies between them. The wrapper method is a combination of learning algorithms and feature subset generation strategies, it evaluates a candidate feature subset of features dependent on the effectiveness of the applied learning algorithm (Alweshah et al., 2022; H. Liu et al., 2019),

where the feature subset is formed according to the opted search strategy, for example, forward selection or backward selection (Urbanowicz et al., 2018). The hybrid methods combine more than one approach to give results that are more accurate and obtain the best feasible feature subset (Remeseiro & Bolon-Canedo, 2019). For instance, one could remove a set of features with a less computationally expensive filter method and then find the best candidate subset features with a wrapper method.

2.3.4 Modeling Stage

There are three main models utilized to treat the SA problem according to the review of literature, which are: lexicon-based model, ML-based model, and DL-based model (Birjali et al., 2021), which can then be further divided into sub-categories as shown in Figure 2.2. A brief description of the main three models is given below:

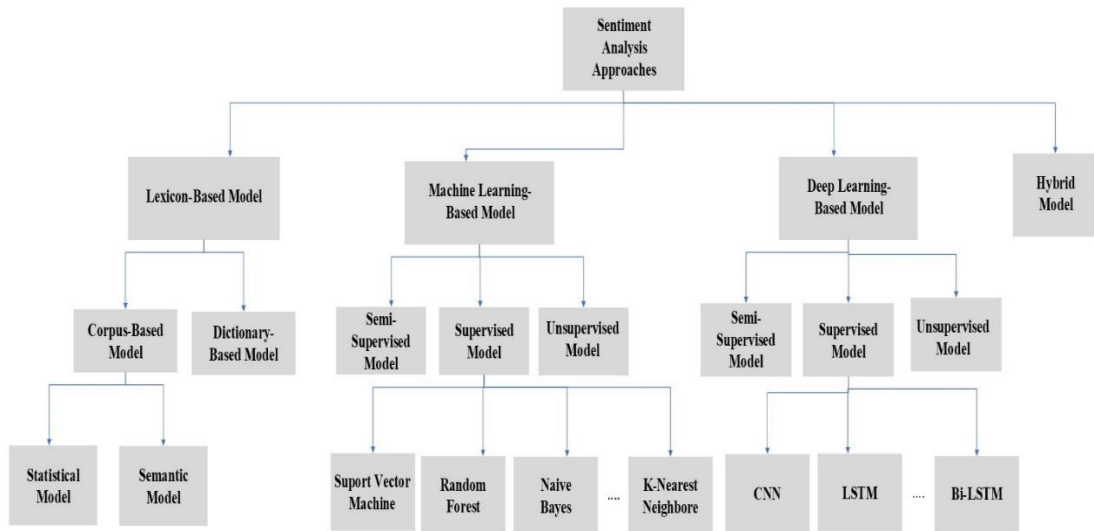


Figure 2.2 Sentiment Analysis Models.

2.3.4(a) Lexicon-Based Model

The lexicon-based model is considered an unsupervised model (Mukhtar et al., 2018). It is a list of lexical features labeled according to their sentiment polarity positive or negative (S. Kumar et al., 2019), which can be separated into two main types:

Dictionary-based and Corpus-based (Assiri et al., 2018; Neha Gupta & Rashmi Agrawal, 2020). The general stages of this model are as follows. Data are extracted from the text reviews. Then pre-processing stages are applied to the extracted data. Finally, the lexicon is applied to the processed data to determine the polarity, as shown in Figure 2.3.

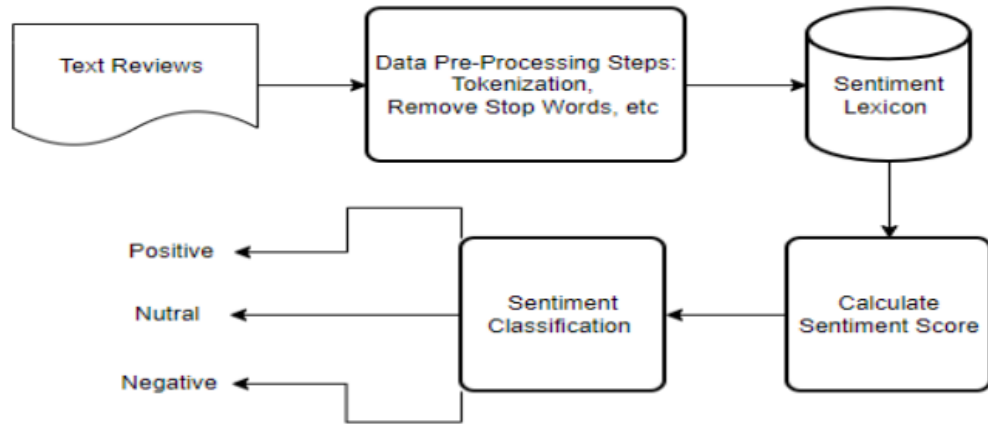


Figure 2.3 Flowchart for Lexicon-Based Model.

This approach works by analyzing how strongly a word is clearly biased toward positive, negative, or neutral meaning. In the literature, there are many lexicons available, such as SentiWordNet, SenticNet, VADER and AFINN (H. Li et al., 2022; Taj et al., 2019). The lexicon-based approaches were applied to a set of languages, for example, Arabic language (Assiri et al., 2018), Hindi language (Kulkarni & Rodd, 2022), and a mix of English and Malay languages (Imanina Zabha et al., 2019).

To get successful results, a lexicon-based model relies on the structure of the language and the relationships between words. The common factor in lexicon models is the way the polarity is determined. Firstly, the polarity of words is determined by assigning 1, 0, or -1 scores to the positive, neutral, or negative words, respectively. Then, all polarities of the words are summed to identify the polarity of the sentence, and then the polarity of the whole document is computed by summing the polarities of its

constituent sentences. However, the quality of the lexicon structure utilized has a significant impact on the accuracy of the findings, where there is no standard structure for lexicons to deal with SA.

These models can be constructed manually or automatically. Although manually constructed lexicons are clearly more accurate than automatically constructed ones, they are a time-and-human resource-consuming procedure. As a result, many studies rely on automatic lexicons, which are constructed utilizing semantic relations of thesauri, morphological techniques, or co-occurrence algorithms in large corpora (Catelli et al., 2022). In the dictionary-based model, a small set of words is picked and subsequently expanded with the help of information from dictionaries. This method's main proposition is that if a word has sentiment polarity, then its synonyms and antonyms do as well. As a result, a new, more complete, set of opinion words can be created from the initial set of words. In this model, computational methods are performed on a text to determine how the text affects the reader's emotions. The performance of a dictionary-based model is strongly dependent on the dictionary chosen. As a result, caution must be exercised before it is decided whether the dictionary is acceptable for the dataset and intended purposes. The second model is looking for rules and patterns in the texts. The co-occurrence statistics of several words are compared to determine word polarity. Its methods rely on patterns that occur together, as well as a seed list of opinion words. Finally, the word orientation is used to categorize reviews.

The main advantage of the lexicon-based model is that it does not require any training or labeling data, making the learning process more simplified and faster. By contrast, the limitation of the dictionary-based model is domain dependence. Extracting

domain-specific opinions using it is not possible. Given that words can have multiple meanings, a positive word in one domain might be a negative word in another. To avoid this limitation a corpus-based model is used. However, the corpus-based model is not as efficient as the dictionary-based model due to the need for a huge corpus to cover all terms, which is a challenging effort.

It is concluded from previous studies that a lexicon-based model relies on the language structure and relationships between words. It is based on determining how distinctly a word is skewed toward positive or negative significance. Furthermore, to provide satisfactory results, the issues of negations, intensifiers, and context-dependent words must be considered. As far as is known, no general structure has been developed for lexicons to deal with SA of languages.

2.3.4(b) Machine Learning-Based Model

Three types of ML models are commonly used for SA: supervised, unsupervised, and semi-supervised models (Ligthart et al., 2021). In contrast to the lexical approach, which does not require time to train the data, the supervised and semi-supervised ML approach does. The supervised approaches are the most common approaches in SA. As a result, the previous studies of this approach will mainly be discussed in this study. Accordingly, in ML models, data is separated into training and testing datasets. Furthermore, these models required a set of processing stages to extract and select the more relevant features before training the classifier. The use of powerful FE and FS techniques can help these models perform well. Figure 2.4 shows a flowchart for a typical ML model that can be used for SA.

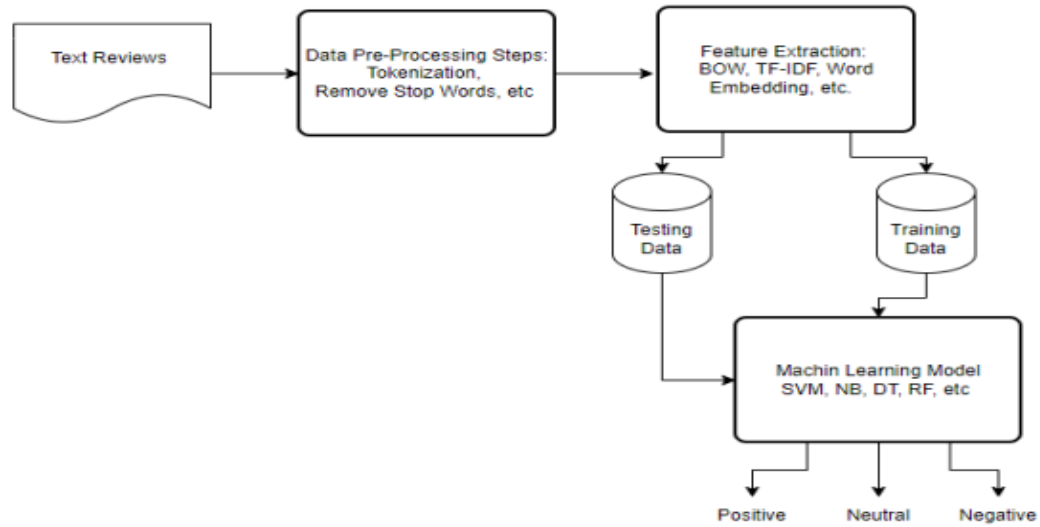


Figure 2.4 Flowchart for Machine Learning-Based Model

The researchers used a variety of supervised ML approaches and diverse ways to extract and represent data. Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are examples of supervised classification algorithms often used in these approaches (Baid et al., 2017a; Naresh & Venkata Krishna, 2021; Raza et al., 2019). Furthermore, several computational linguistic methods are utilized to encode Twitter reviews as information vectors, which are deemed an appropriate format for these approaches enabling them to deal with the extracted features data, such as BOW, TF, and TF-IDF (S. & K.V., 2020).

Regarding the literature studies, SVM, NB, RF, and KNN are the most broadly used classification techniques in the ML supervised models. SVM is estimated to be the most efficient of these techniques. However, from the earlier studies, the conclusion is drawn that to achieve a good result using ML models, suitable FE and FS methods need to be employed.