

# **SENTIMENT ANALYSIS ON SOCIAL MEDIA USING SOFT COMPUTING TECHNIQUES**

A THESIS

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY

FOR THE AWARD OF THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

IN

**COMPUTER ENGINEERING**

SUBMITTED BY

**ARUNIMA JAISWAL**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
DELHI- 110042 (INDIA)**

**2020**

# **SENTIMENT ANALYSIS ON SOCIAL MEDIA USING SOFT COMPUTING TECHNIQUES**

BY

**ARUNIMA JAISWAL**

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY IN PARTIAL FULFILMENT  
OF THE REQUIREMENTS FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

IN

**COMPUTER ENGINEERING**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**DELHI TECHNOLOGICAL UNIVERSITY**

**(Formerly Delhi College of Engineering)**

**DELHI- 110042 (INDIA)**

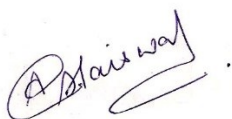
**2020**

©DELHI TECHNOLOGICAL UNIVERSITY-2020  
ALL RIGHTS RESERVED

## **CANDIDATE DECLARATION**

I hereby certify that the research work which is being presented in this thesis entitled "**Sentiment Analysis on Social Media using Soft Computing Techniques**" in fulfilment of requirements of the award of degree of Doctor of Philosophy, is an authentic record of my own research work carried out under the supervision of Dr. Akshi Kumar.

The matter presented in this thesis has not been submitted elsewhere in part or fully to any other University or Institute for award of any degree.



**Arunima Jaiswal**

2K16/PHD/CO/02  
Department of Computer Science & Engineering  
Delhi Technological University  
Delhi, India-110042.



# DELHI TECHNOLOGICAL UNIVERSITY

(Govt. of National Capital Territory of Delhi)

BAWANA ROAD, DELHI – 110042

## CERTIFICATE

Date: 23-10-2020

This is to certify that the thesis entitled “**Sentiment Analysis on Social Media using Soft Computing Techniques**” done by Arunima Jaiswal, Roll no. 2K16/PHD/CO/02 in Department of Computer Science & Engineering, Delhi Technological University is an authentic work carried out by her under my guidance.

This work is based on original research and the matter embodied in this report has not been submitted earlier for the award of any degree or diploma to the best of my knowledge and belief.

**Supervisor**

---

**Dr. Akshi Kumar**

Assistant Professor

Department of CSE, DTU

## ACKNOWLEDGEMENT

I feel pride in placing on record my deep gratitude to my Ph.D. supervisor, Dr. Akshi Kumar, Assistant Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, who guided me throughout this long journey of four years of research work. Whether it was review of literature or clearing doubts, she spared quality time to guide me despite her busy schedule. She not only helped in solving all problem with comfortable ease but also encouraged and motivated me to sail through in difficult times. I have gained immensely from her comments and suggestions regarding the technical quality of the work. It is hard to imagine successful completion of such a research work without her guidance and care.

I am very grateful to Prof. Rajni Jindal, Head, Department of Computer Science & Engineering, Delhi Technological University, Delhi, for her constant encouragement and support to accomplish this task.

I am thankful to all the other faculty members of Department of Computer Science and Engineering, Delhi Technological University, Delhi, for the motivation and inspiration. I am also thankful to all non-teaching staff of CSE Department, DTU who have helped me directly or indirectly in completion of this research work.

I wish to pay high regards to my parents and to my parent in-laws for their invaluable support, best wishes and encouragements. I am grateful to my husband and child for their love, patience and support for completing this project. Without them this research would not have been possible.



Arunima Jaiswal

2K16/PHD/CO/02 (PhD Scholar)

Department of Computer Science & Engineering

E-mail: [arunimajaiswal@gmail.com](mailto:arunimajaiswal@gmail.com)

## ABSTRACT

*“The analysis of variance is not a mathematical theorem, but rather a convenient method of arranging arithmetic”* ----Ronald Fisher

Social media can be described as the VUCCA world that is Volatile-Uncertain-Complex-Chaotic-Ambiguous that generates enormous amount of online user content, which can further be examined to get insights for social intelligence. Undeniably, with the quantum of opinionated data on social media, sentiment analysis now finds use in various marketing, business and government applications. But the noise, high-dimensionality, imbalance, heterogeneity, multimodality and multi-linguality associated with the social media data makes the task of sentiment analysis challenging. Further, the growing use of micro-texts (creative spellings, slangs etc.) compounds the linguistic challenges of sentiment analysis. Good features are considered as the backbone for any learning model, and good feature creation often needs adequate domain knowledge, creativity and time. This necessitates examining new computational methodologies for finding optimal feature set which improves the performance of the sentiment classifier in terms of predictive accuracy and result comprehensibility. One such consortium of techniques is referred to as soft computing, which provides robust and low cost solutions that could cater well with these upshots. In this research, we examine sentiments using soft computing on benchmark (SemEval 2016 & 2017, Sentiment140, IMDb movie review corpus) and scrap (textual topic based) data from social media namely, Twitter, Tumblr, etc. Experiments for sentiment analysis using TF-IDF are conducted on these datasets using ensemble (random forests, bagging, boosting, gradient boosting, stochastic gradient boosting and extra trees) and baseline machine learning techniques (naive bayesian, support vector machine, multilayer perceptron, decision tree and k- nearest neighbour). This is followed by the application of swarm intelligence techniques (namely particle swarm, binary grey wolf, binary moth flame) for feature optimization on Twitter (benchmark) datasets for enhanced textual sentiment analysis. Also, in this study, catering to the challenge of selecting the essential features each time, which is altogether a computationally hard task, deep convolution neural network using GloVe is used which automatically learns features at multiple level of abstraction without depending completely on hand-crafted features. Deep learning techniques have hierarchical learning capabilities and at the same time, the use of adaptive and heuristic optimization to select a near-optimal set of input variables that would minimize variance and maximize generalizability of the learning model, is highly desirable to achieve high prediction accuracy. Based on this, we finally propose a cognition-driven model for sentiment classification which is built on the concord of deep learning (convolution neural network), swarm optimized machine learning (wolf-search algorithm and decision tree). All the results are evaluated using accuracy, precision & recall. The proposed model compares favourably to state-of-the-art approaches and achieves an average performance accuracy of 89.5 on SemEval 2016 & 2017 datasets.

# Table of contents

<i>Candidate declaration</i>	<i>i</i>
<i>Certificate</i>	<i>ii</i>
<i>Acknowledgement</i>	<i>iii</i>
<i>Abstract</i>	<i>iv</i>
<i>Table of contents</i>	<i>v</i>
<i>List of abbreviations</i>	<i>viii</i>
<i>List of figures</i>	<i>xi</i>
<i>List of tables</i>	<i>xiii</i>
<i>List of algorithms</i>	<i>xiv</i>
<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1 Introduction.....	2
1.1.1 Motivation.....	5
1.1.2 Social Media.....	7
1.1.3 Sentiment Analysis.....	9
1.1.4 Soft Computing Techniques.....	10
1.2 Problem Statement and Research Objectives.....	14
1.3 Significance of Sentiment Analysis on Social Media using Soft Computing....	14
1.4 Organization of thesis.....	16
1.5 Chapter summary.....	17
<b>Chapter 2 Systematic literature review .....</b>	<b>18</b>
2.1 Formulation of Research Questions.....	20
2.2 Search Strategy.....	20
2.3 Study Selection.....	21



2.4	Quality Assessment.....	22
2.5	Data Extraction.....	22
2.6	Data Synthesis.....	22
2.7	Literature Survey.....	33
2.8	Key Observations and Research Gaps.....	39
2.9	Chapter summary.....	40

**Chapter 3 Sentiment Analysis using Soft Computing: Convergence with Web 2.0.....41**

3.1	Research Objective 1.....	43
3.2	Methodology.....	43
3.3	Findings.....	44
3.4	Chapter summary .....	49

**Chapter 4 Sentiment Analysis using Machine Learning, Swarm based Learning and Deep Learning.....50**

4.1	Research Objective 2.....	51
4.2	Methodology.....	52
4.2.1	Application of Baseline Supervised Machine Learning .....	52
4.2.2	Application of Swarm Intelligence and Machine Learning techniques.....	52
4.2.2.1	Application of Binary Grey Wolf.....	54
4.2.2.2	Application of Binary Moth Flame.....	60
4.2.2.3	Application of Particle Swarm Optimization.....	65
4.2.3	Application of Deep Learning Technique.....	67
4.3	Findings.....	67
4.3.1	Findings of Methodology 1.....	67
4.3.1.1	Findings for Twitter and Tumblr Datasets.....	67
4.3.1.2	Findings for SemEval 2016, 2017 Datasets.....	72
4.3.1.3	Findings for IMDb movie review & Sentiment140 Datasets.....	76
4.3.2	Findings of Methodology 2.....	78
4.3.2.1	Findings of BGW and BMF for SemEval 2017 Dataset.....	78
4.3.2.2	Findings of BGW and BMF for SemEval 2016 Dataset.....	80

4.3.2.3	Findings of PSO for SemEval 2016, 2017 Datasets.....	83
4.3.3	Findings of Methodology 3.....	85
4.4	Chapter summary.....	87
 <b>Chapter 5 Novel Framework for Sentiment Analysis using Soft Computing.....</b>		<b>89</b>
5.1	Research Objective 3.....	91
5.2	Methodology.....	91
5.2.1	Convolution Neural Network (CNN).....	94
5.2.2	Decision Trees (DT).....	95
5.2.3	Meta-heuristic Optimization using Wolf Search Algorithm.....	95
5.2.3.1	Merging with other wolves.....	97
5.2.3.2	Preying.....	98
5.3	Findings.....	99
5.3.1	Performance of the proposed CNN + $WSA_{DT}$ .....	99
5.3.2	Comparison of DT with other ML techniques.....	100
5.3.3	Comparison of WSA with other meta-heuristic optimization algorithms.....	102
5.3	Chapter Summary.....	103
 <b>Chapter 6 Conclusion and Future Scope .....</b>		<b>104</b>
6.1	Research summary.....	105
6.2	Limitation of study.....	107
6.2	Future directions.....	107
6.3	Conclusion.....	107
 <b>References .....</b>		<b>112</b>
 <b>Appendix-A List of publications .....</b>		<b>119</b>

## LIST OF ABBREVIATIONS

<b>A</b>	Accuracy
<b>AI</b>	Artificial Intelligence
<b>AP</b>	Average Precision
<b>AUC</b>	Area under the Curve
<b>BGW</b>	Binary Grey Wolf
<b>BMF</b>	Binary Moth Flame
<b>Bos</b>	Boosting
<b>Bgg</b>	Bagging
<b>BCS</b>	Binary Cuckoo Search
<b>Cf</b>	Confidence
<b>CK</b>	Cohen's Kappa
<b>CNN</b>	Convolution Neural Network
<b>CM</b>	Confusion Matrix
<b>DNN</b>	Deep Neural Network
<b>DL</b>	Deep Learning
<b>DT</b>	Decision Tree
<b>EC</b>	Evolutionary Computing
<b>EER</b>	Equal Error Rate
<b>EM</b>	Ensemble Methods
<b>ER</b>	Error Rate
<b>ET</b>	Extra Trees
<b>F</b>	F score

<b>FAR</b>	False Acceptance Rate
<b>FC</b>	Five Cross
<b>FL</b>	Fuzzy Logic
<b>FP</b>	False Positive
<b>FN</b>	False Negative
<b>FRR</b>	False Rejection Rate
<b>GA</b>	Genetic Algorithm
<b>GB</b>	Gradient Boosting
<b>GM</b>	Geometric Mean
<b>k-NN</b>	K-Nearest Neighbors
<b>LR</b>	Linear Regression
<b>LogR</b>	Logistic Regression
<b>LSTM</b>	Long Short Term Memory
<b>MiAF</b>	Micro Average F Score
<b>MAF</b>	Macro F1 Score
<b>MAER</b>	Macro Average Error Rate
<b>ML</b>	Machine Learning
<b>MLP</b>	Multilayer Perceptron
<b>MR</b>	Multiple Regression
<b>NB</b>	Naïve Bayesian
<b>NIA</b>	Nature Inspired Algorithms
<b>NMAE</b>	Normalized Mean Absolute Error
<b>NN</b>	Neural Networks
<b>NLP</b>	Natural Language Processing
<b>P</b>	Precision
<b>PSO</b>	Particle Swarm Optimization
<b>R</b>	Recall

<b>RF</b>	Random Forests
<b>RNN</b>	Recurrent Neural Network
<b>RQ</b>	Research Question
<b>SA</b>	Sentiment Analysis
<b>SC</b>	Soft Computing
<b>SI</b>	Swarm Intelligence
<b>SGB</b>	Stochastic Gradient Boosting
<b>SLR</b>	Systematic Literature Review
<b>Su</b>	Support
<b>Sp</b>	Specificity
<b>Sn</b>	Sensitivity
<b>SVM</b>	Support Vector Machine
<b>TP</b>	True Positive
<b>TN</b>	True Negative
<b>TC</b>	Ten Cross
<b>TF IDF</b>	Term Frequency–Inverse Document Frequency
<b>TWC</b>	Twenty Cross
<b>WSA</b>	Wolf Search Algorithm

# LIST OF FIGURE(S)

Figure 1.1. Data generation using IoT, smart phones and smart devices.....	4
Figure 1.2. Relevance of sentiment analysis in building smart city.....	5
Figure 1.3. Generic SA process.....	9
Figure 1.4. Types of SC Techniques.....	12
Figure 1.5. Relation between SC, ML and DL.....	13
Figure 2.1. Phases of SLR.....	19
Figure 2.2. Review Process.....	22
Figure 3.1. Various domains covered for SA in past decade.....	44
Figure 3.2. Distribution of different benchmark datasets.....	45
Figure 3.3. Year-wise cumulative assessment of random tweets.....	46
Figure 3.4. Year-wise distribution of number of papers.....	46
Figure 3.5. Distribution of papers after year 2014.....	47
Figure 3.6. Quantitative extent of use of SC techniques over past decade.....	48
Figure 3.7. Distribution of SC techniques over the past decade.....	48
Figure 4.1. Systematic workflow for optimized sentiment classification.....	54
Figure 4.2. BGW hierarchy (dominance increases from bottom up).....	55
Figure 4.3. Hunting mechanism of Grey Wolves.....	56
Figure 4.4. Transverse navigation mechanism of moths in moonlight.....	60
Figure 4.5. Spiral flight trajectory around close luminous sources.....	60
Figure 4.6. The concept of a flying particle.....	66
Figure 4.7. Accuracy results on using ML techniques on Twitter, Tumblr .....	71

Figure 4.8. Accuracy results on using ML techniques on SemEval 2016, 2017.....	73
Figure 4.9. Accuracy results on using Ensemble techniques on SemEval 2016, 2017.....	75
Figure 4.10. Accuracy results on Sentiment140 and IMDb movie reviews.....	76
Figure 4.11. Features selected using BGW and BMF in SemEval 2017 .....	79
Figure 4.12. Accuracy results on using SI+ML on SemEval 2017.....	80
Figure 4.13. Features selected using BGW and BMF in SemEval 2016.....	82
Figure 4.14. Accuracy results on using SI+ML on SemEval 2016.....	82
Figure 4.15. Accuracy results on using SI (PSO) + ML on SemEval 2016, 2017.....	84
Figure 4.16. Architecture of CNN model.....	86
Figure 5.1. Architecture of proposed CNN- <sub>WSA</sub> DT model.....	92
Figure 5.2. Convolution operation.....	94
Figure 5.3. WSA in action.....	96
Figure 5.4. ROC for SemEval 2016 (DS-I) and SemEval 2017 (DS-II).....	100
Figure 5.5. Comparison of supervised learning techniques using accuracy.....	101
Figure 5.6. Accuracy (%) of DT with and without WSA optimization.....	102

## LIST OF TABLES(S)

Table 2.1: Summary of the studies undertaken for review.....	23
Table 3.1: Mapping of the techniques with varied domains.....	47
Table 4.1: Details about SC techniques used.....	68
Table 4.2: Summary of results for ‘Rio Olympics’.....	69
Table 4.3: Summary of results for ‘Release of Pokemon Go Gen’.....	70
Table 4.4: Summary of results for ‘US presidential elections’.....	70
Table 4.5: Summary of results for ‘Donald Trump’s claims of Muslim Ban’.....	70
Table 4.6: Description of ensemble techniques.....	74
Table 4.7: Summary of results for ensemble techniques on SemEval 2016, 2017.....	75
Table 4.8: Accuracy gain using BGW and BMF in SemEval 2017.....	78
Table 4.9: Percentage of features selected using BGW and BMF in SemEval 2017.....	79
Table 4.10: Result summary for SemEval 2017.....	80
Table 4.11: Accuracy gain using BGW and BMF in SemEval 2016.....	81
Table 4.12: Result summary for SemEval 2016.....	83
Table 4.13: Classifier accuracy before & after feature-selection for SemEval 2016.....	83
Table 4.14: Classifier accuracy before & after feature-selection for SemEval 2017.....	84
Table 4.15: Results using CNN (DL) on SemEval 2016, 2017.....	86
Table 5.1: Results of CNN + $w_{sADT}$ .....	99
Table 5.2: Comparative analysis of CNN and hybrid model for SemEval 2016, 2017.....	99
Table 5.3: Feature selection using TF-IDF+WSA.....	101
Table 5.4: Accuracy comparison of optimization techniques.....	102
Table 6.1: Mapping of research objectives with the list of publications.....	105



# LIST OF ALGORITHMS

Algorithm 1: GWO.....	57
Algorithm 2: BGW.....	59
Algorithm 3: MFO.....	61
Algorithm 4: BMF.....	64
Algorithm 5: PSO.....	66
Algorithm 6: Proposed model (CNN+ $w_{SA}DT$ ).....	93
Algorithm 7: WSA.....	96

# Chapter 1

---

## *Introduction*

# Chapter 1

## Introduction

This chapter briefly introduces the fundamental concepts related to research area. Section 1.1 discusses the motivation and challenges of the chosen research area and provides a brief description of the key terminologies namely, sentiment analysis, social media and soft computing. Section 1.2 formulates the statement of research problem and decomposes the unified research question, sub questions leading towards certain identified research objectives. Section 1.3 provides a brief description notifying the significance of sentiment analysis using soft computing techniques. Further, section 1.4 comprises of organization of thesis. Section 1.5 discusses the summary of the chapter.

### 1.1 Introduction

The incessantly evolving dynamics of the Web in terms of the volume, velocity and variety of opinion-rich information accessible online, has made research in the domain of Sentiment Analysis (SA) a trend for many practical applications which facilitate decision support and deliver targeted information to domain analysts. Interestingly, the buzzing term 'big data' which is estimated to be 90% unstructured [1] further makes it crucial to tap and analyse information using contemporary tools. Text mining models define the process to transform and substitute this unstructured data into a structured one for knowledge discovery. Use of classification algorithms to intelligently mine text has been studied extensively across literature [2]. SA is defined as the computational study of people's opinions/attitudes/emotions towards an entity [3-5]. It offers a technology-based-solution to understand people's reactions, views & opinion polarities (positive/negative/neutral) in textual content available over social-media sources. SA is typically a text classification tasks, where effective feature selection plays a key role in determining the sentiment classification accuracy [6]. Feature selection is one of the most important & indispensable process of classification in many real-world applications. A minor error in the process of classification can lead to substantial impact on information processing in varied fields like disease detection (COVID) in medical science, customer identification & authentication for online-banking etc. Subsequently, it is essentially critical to have a precise classifier with high and predictable exactness which can be applied in real-world applications that can automatically perform feature extraction so that there is no requirement for manual feature extraction.

Also, due to the abundant volume, velocity and variety of opinion rich Web data accessible online via Internet, a significant part of the recent research is concentrating on the ongoing area of text mining field which is called as sentiment analysis [7]. Additionally, because of the humongous volume-velocity-variety of sentiment rich Web data available online through Internet, a noteworthy aspect of the ongoing research is

focusing on the progressing area of text mining field which is called as sentiment analysis [8]. Research studies & pragmatic applications in the field of SA have increased in the previous decade with the change and extension of Web from passive-provider of content to an active socially-aware-distributor of collective intelligence. This new collaborative Web called as Web 2.0 [6] is largely comprising of Web-based technologies such as comments/blogs/wikis, social media portals like Facebook etc. This new Web facilitates building social networks, encouraging broader spectrum of expression, enabling creation of community, sharing of dialogue & knowledge and attracting authentic audiences by modes of varied tools and technologies. This confluence of social media-mobile-analytics-cloud has presented the novel SMAC [6] technology paradigm, which has altered the operative environment & user engagement on Web remarkably. There has been a lateral shift from the conventional e-commerce (electronic commerce) to substantial s-commerce (social commerce) nowadays. S-commerce is regarded as a subset of e-commerce that has upheld nearly all main innovative practices which provide assistance to online commercial activities including retailing and marketing by incorporating social network(s).

Due to this, there is expansion in the scope of commercial activities that enables the users to exploit the forum either by discussing, sharing, analysing, criticising, comparing, appreciating and doing research about various products, brands or services through social platforms like Voonik, Facebook, Twitter etc. This information could be discovered for mutual benefit by the customer and the organization as well. Data analytics on such social web-based corpora has been an ongoing-trend where online texts etc. are transformed into an opinion-rich knowledge-base that can influence efficient decision making. This signifies that the mass is relying on such online user-generated content (UGC) for the opinions that comprehensively marks the increasing importance of SA in our day to day lives. For example, it has often been used by people to express and vent out their opinions, frustrations etc. via social media platforms [8-9] like Facebook, Twitter etc. It has been widely used these days and people are becoming more aware about the current trending topics prevailing across the globe. Using this as a medium, almost every individual is getting involved in it by posting his views and concerns about the prevailing topic. For example, on the Supreme Court's verdict about "teen tallaak", the prime effect of public discourse could very well be understood from these social media platforms. We can say that, it has a powerful impact on the public concerning government policies, regulation proposals, legal matters, law amendments etc.

Also, advancement of Web 2.0 has completely revolutionized the way people communicate and exchange information among them using social media [10] such as Twitter etc. People use smart technology based solutions (including IoT's, sensors etc.) at an unprecedented scale for enhancing their quality of the life and building a smart city. Such cities that involve constant engagement of its citizens and the advanced technology driven services are called as cognitive cities. The cognition in a cognitive city occurs

primarily due to the consumption and generation of data among the whole city i.e. from citizen to citizen and from citizen to the system.

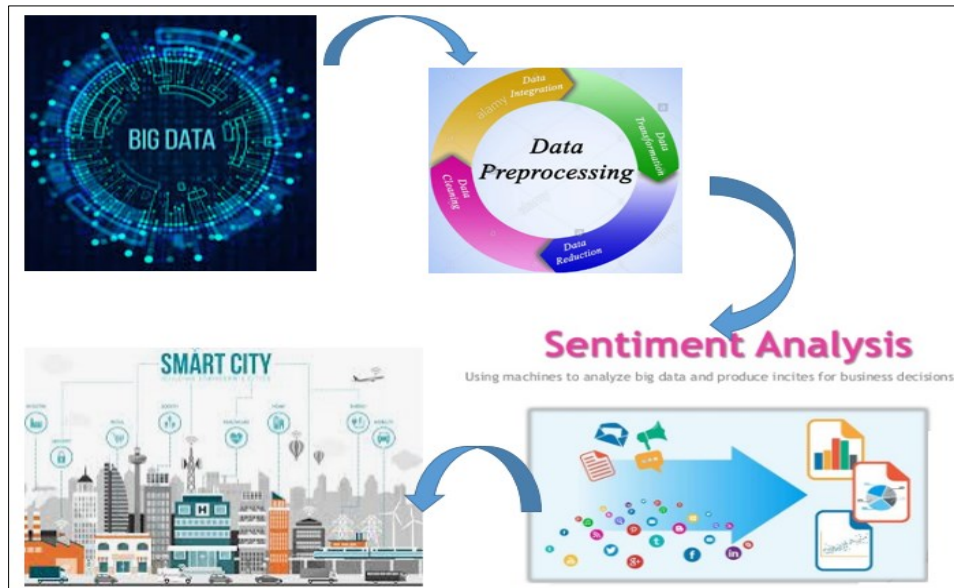
Sentiment analysis [5, 8] serves as info-foundation for cognitive cities as they have the ability to harness the opinions or sentiments accurately based on the computation technology applied for efficient adaptability and improved scalable learning. They become vast source of opinion-rich data that needs to be processed & comprehended well in order to aid improved and smart decision making by the people for any product, service or policy etc.

Adequate and appropriate usage of this sentiment-rich data could be further used for analysis purposes that would help in solving problems related to tourism, environment disaster management, meteorology, policy-making, feedbacks, gaming, entertainment, start-ups, business making, recommendations etc. Amongst all, the common thing is the real-time big data that is being acquired from crowd-sensing, IoTs, smart phones etc. which is often related with the community via social media as shown in figure 1.1.



**Fig 1.1.** Data generation using IoT, smart phones and smart devices

So, we can say that sentiment analysis is going to be a prime aspect for upgrading, augmenting and intensifying smart-city-governance values in order to improve the urban ecosystem, as shown in figure 1.2.



**Fig 1.2.** Relevance of sentiment analysis in building smart city

Hence we can say that Internet has humongous data that needs to be examined meticulously for determining apt sentiments. There have been several ongoing research projects that apply SA [11] to the variety of web based corpora including social media forums such as Twitter, Facebook etc. for efficient decision making. Literature survey in the proposed research area identified it as a potential dynamic direction of research with promising applications and technologies.

### 1.1.1 Motivation

SA has become an integral part of social media [7] and has become indispensable in appropriate decision making. It is a precise strategy for predicting accurate opinions related to the users. For example, if one wants to buy a mobile phone, instead of relying on the friends and relatives viewpoints, he will find gamut of information available about the product's (mobile's) reviews on many public forums which could be beneficial to him in his decision making of buying a good mobile phone as per his preferences. Opinion mining (often taken as synonymous to SA) is a computational identification and classification of opinions that are being expressed as written text on any subject matter of interest [5-6, 8-9]. Till now, there have been various issues prevailing in this field, such as polarity based classification of the sentiments, classifying sentiments based on feature assessment, developing a novel approach for classifying sentiments for any specific domain utilizing supervised or unsupervised learning procedures or lexicon based or hybrid methods. SA is almost used in all major areas where enormous text documents can be processed for obtaining respective sentiments attached to them and to track different survey responses or customer reviews or product reviews etc. for appropriate decision making. SA has emerged as a notable direction of research with scientific trials and promising applications being explored substantially. It has turned out as an exciting

new trend with a gamut of practical applications that range from business intelligence; finance domain; political domain; entertainment domain; corporate management, amongst others.

Thus, we can say that sentiment detection & classification is the cutting-edge fad for social media analytics on Web. Covering wide spectrum of applications related to healthcare, education, finance, tourism, media, consumer markets, stock exchange and government etc., distilling the voice of people to gain insight to target information and reviews is non-trivial. Social media can be described as the VUCCA world that is **Volatile-Uncertain-Complex-Chaotic-Ambiguous** [1, 11] that generates enormous amount of online user content (UGC), which can further be examined for getting proper insights for market, business or government intelligence etc. Artificial Intelligence (AI) technologies for instance machine learning and natural language processing (NLP) facilitate contextual understanding and allow personalization of products and services for online customers. Intelligent adaptive models are required to deal with the information overload on the chaotic and complex social media portals and to fully realize the benefits of social media for a connected, optimized, transparent and proactive marketplace. Certainly, user-generated big-data is a substantial source for enhancing business competitiveness. Data classification is a promising analytic technique which is extensively used to solve the IoT and big-data centred problems for various business or personal objectives. Recently, sentiment analysis or emotion AI has been used to determine insights pertaining to a topic, brand or event [11]. It is the use of natural language processing (NLP) and computational linguistics to interpret and classify online conversations in terms of positive and negative mentions facilitating decision making [10].

Social media is an informal mode of communication which comprises of immense usage of slangs, short forms, mal-formed words, mash-up words and colloquial expressions. At the same time, most real-time datasets scrapped from social media suffer from imbalance class distribution (skewed data), that is, the dataset is imbalanced. This augments the uncertainty and imprecision within the presented social web content (UGC). Sentiment analysis (SA) is typically a text classification task [5] which relies on converting the natural man-made language to a form of representations (features) that are easily interpreted and learned by machines. In general the features can be divided into two broad categories, namely meta-features, such as stop word counts, word counts, punctuation counts, the language of text, the length of characters etc. and text-based features, such as features extracted using tokenization, vectorization, stemming, part-of-speech tagging, and the named entity recognition. Analysing and classifying such colossal, noisy, uncertain social web data (UGC) affects the quality of sentiments derived and presents novel challenges to feature selection [9]. One of the non-trivial and challenging sub-tasks of SA is therefore feature engineering. Feature engineering primarily uses domain knowledge for extracting features in order to reduce data complexity and create patterns for learning algorithms to work. This process is arduous and expensive in terms

of time and competence. Undoubtedly, it has a significant role in enhancing the sentiment classification accuracy.

It is quite evident from our pertinent literature that optimal feature selection primarily relies on improving the classifier performance, reducing the dimensionality, removing noise and helping in visualizing the data for appropriate model selection. As it is said that good ingredients make up the good recipes, likewise, good features serves as the backbone for any machine learning model, and good feature creation majorly requires proper domain knowledge, creativity, and lots of time. Thus, in order to boost the predictive sentiment accuracy of learned models and to enhance the learning efficiency with respect to reduce storage needs and computational costs, it is quite imperative to construct an intelligent feature selection model. This fosters the need to search for enhanced and optimized feature engineering techniques that could accurately and efficiently classify sentiments. One such consortium of techniques that can significantly improve feature extraction and consequently sentiment classification is soft computing (SC). These techniques have the ability to capture imprecision, uncertainty and the dimensionality in user-generated social media content [12-14]. These aid in handling and modeling the complexity associated with the real-world problems providing robust & low cost solutions. Feature selection, noise removal and parameter optimization are few problems which can be solved using SC techniques for improved and enhanced SA.

SC techniques also merges novel computational techniques that often mimic consciousness and cognizance in several important aspects. Application of SC techniques for SA on social media is thus a promising direction of research covering almost all the practical domains for discovering, exploring and understanding the extensibility of man-made expressions [13].

### **1.1.2 Social Media**

SA is an imminent and recent area of research where user-generated (UGC) real-time data obtained via crowd-sensing & smart devices is continually examined for getting proper visions leading to better performances. Social media denotes websites, applications etc. that are designed to permit people to share all kinds of multimodal content quickly, efficiently, and in real-time, thus, generates high volume of user-generated data. People often choose expressing and voicing their emotions & opinions over major social media channels for example, blogs, review websites, posts and micro-blogs. The usage of these social media platforms has observed an explosive growth in last decade where people are connected to each other via links or connections etc. where they are allowed to share or posts without any geographical boundaries with the help of their computing devices such as mobiles, ipods, laptops etc. Data obtained from such social micro-blogs is majorly voluminous and varied. People express their sentiments or opinions over these social media networks making it a huge 'sentiment-rich corpuses from which strategic data can



be analysed for efficient decision making [15-18]. A large portion of big data comprises of user-generated content which is also called as UGC. It is largely created by the general public or consumers instead of by any marketing professionals. With the expansion of the Internet and social media, a vast amount of UGC is posted in textual or other formats (images, audios, videos), such as tweets of Twitter, videos of YouTube and product reviews of Amazon etc. This has indeed transformed the operational environment and user engagements on Web remarkably. SA on social media such as Twitter, Tumblr, Facebook etc. has been a research trend with persistent and sustained studies on it that intend to improve & optimize the accuracy of the results obtained for SA. Henceforth, a constant need to leverage this UGC for social media analytics has been recognized by both researchers & practitioners as the current need of the hour.

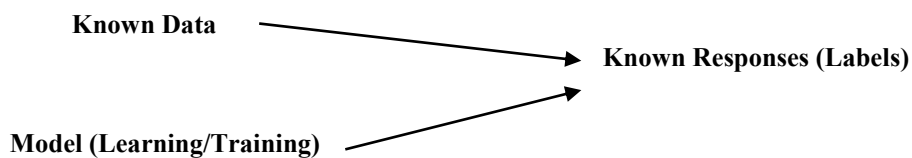
Amongst the Web 2.0 tools, Twitter [8] has evolved as a major revolution in field of social media and has a global reach. It has been the most preferred social channel from which sentiment rich data can be extracted. Originally launched in 2006, Twitter is the currently the most popular and impactful micro-blogging service connecting millions of people worldwide. It is a freely available social networking microblogging media service where a registered user is allowed to broadcast short messages or posts called “tweet” to other registered users in real-time [10]. SA on Twitter has gained popularity due to intrinsic characteristics of the real-time messages shared on it. This is primarily due to the fact that the post size characterizes short text with character-set limit of 280. It has various applications covering wide spectra of domains and has eventually become an indispensable part of an individual's daily digital routine life. Moreover, due to its global connectivity with the diverse user-base and active participation from the users makes it the qualitative and quantitative base for analysing sentiments [19]. Utility and usefulness of Twitter is escalating day by day. Majority of the famous personalities including actors, politicians, sportspersons etc. are having their Twitter accounts and public follow them. For example, on being selected as the 45<sup>th</sup> US president, our honourable Prime Minister Mr. N. Modi tweeted on his wall by congratulating honourable President Mr. D. Trump for his clear, open and unbiased winning. He even tweeted him about improving the Indo-US relationship. Other example of tweet analysis focused specifically about the features of ‘iphone6’ and mixed opinions were observed for the same. Twitter has then eventually evolved as a huge source of various kind of data. People post their real-time tweets about their opinions on any topic of their interest, talking about the current issues prevailing in the world, discussing about their positive or negative sentiments about the products, services they may be using in their routine life. Such tweeting has effectively made the organizations to survey these portals for knowing the rating of their products [20], services etc. so that based on the sentiments of the users they can meticulously make the desired changes to their products or services for further improvement and increasing business strategies. Eventually, there occurs a need to focus on these tweets and consequently classifying them into positive, negative or neutral sentiments. The main ideology is how to maximize the utilization of such tweets or data and modulate them into our research procedures for carrying out future endorsements. Tumblr is another

micro-blogging [21] portal, which came almost around the same time as Twitter [22] but has gained popularity recently due to some value-added features such as posting images, audios, videos, and other media depending on user's knowledge for customizing, managing and uploading such files to create tumblogs (short-form blogs). 'Tumblogging' has not been used much in research studies whereas 'Tweeting' has been the core of most prominent baseline studies.

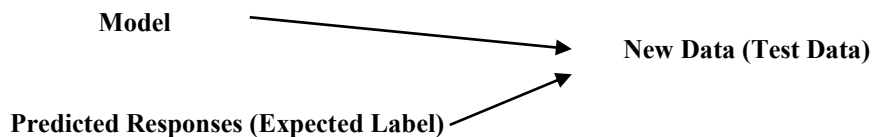
### 1.1.3 Sentiment Analysis

SA has been established as a typical text classification task across pertinent literature. A classification task is an instance of supervised learning from examples. In a supervised learning model the data (observations, measurements, etc.) are labelled with pre-defined classes and the test data are classified into these classes too. It is a two-step process:

- *Learning (training)*: Learn a model using the training data



- *Testing*: Test the model using unseen test data to assess the model accuracy



The process of SA is depicted in the figure 1.3 below:

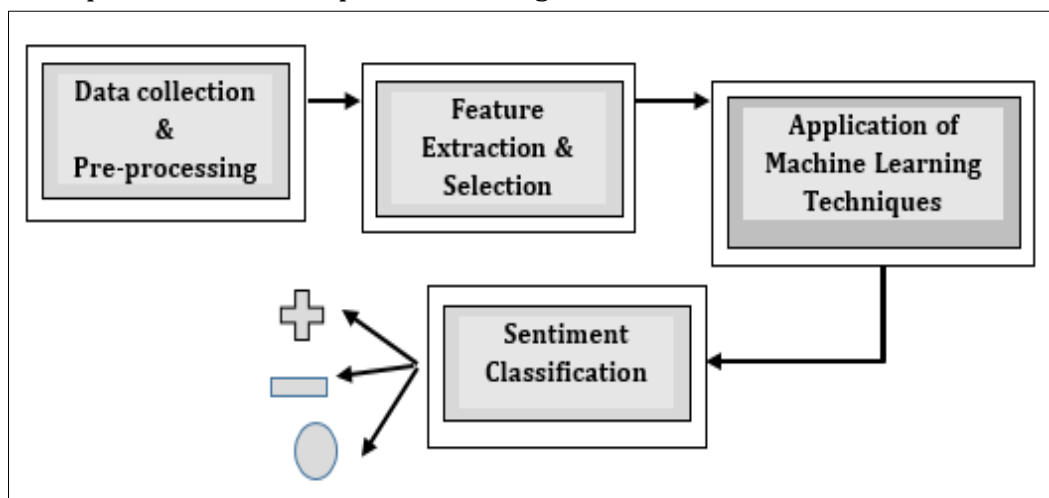


Fig 1.3. Generic SA process

First step is data collection where data is gathered from any social media forums. The collected data is further cleaned in order to remove noise and irregularities present in the data through data pre-processing. Next step is feature extraction and selection. Extraction is obtaining valuable features from the existing data. And selection is selecting a subset from the original pool of features. Last step is application of the selected ML models in order to classify sentiments efficiently.

There has been rapid growth of research in the field of SA and it has emerged as an upcoming and active area of research. It has spread to almost all spheres including computer science, social science, and management sciences etc. It has observed its applicability in majorly all the domains including business and social media like news forum, political debates, reviews, blogs, Facebook, microblogging ( for example Twitter) etc. where enormous opinionated data is available in digital formats. Recently, it is also seen that SA has started covering industrial applications as well, thereby, covering all the spectrum of domains [15-17].

#### **1.1.4 Soft Computing Techniques**

As discussed, generic SA [1, 4, 10-11] task includes Data collection; feature selection; sentiment classification and sentiment polarity detection. Effective feature selection is a computationally hard task and has a significant role in determining the sentiment classification accuracy. Moreover, the increased dimensionality, complexity and fuzziness in the user-generated Twitter data further fosters the need to look for improved and optimized sentiment classification techniques. Studies are constantly being conducted to explore new paradigms which handle uncertainty, imprecision, approximation, partial truth, fuzziness and allow replication of human intelligence for personalized and tractable results.

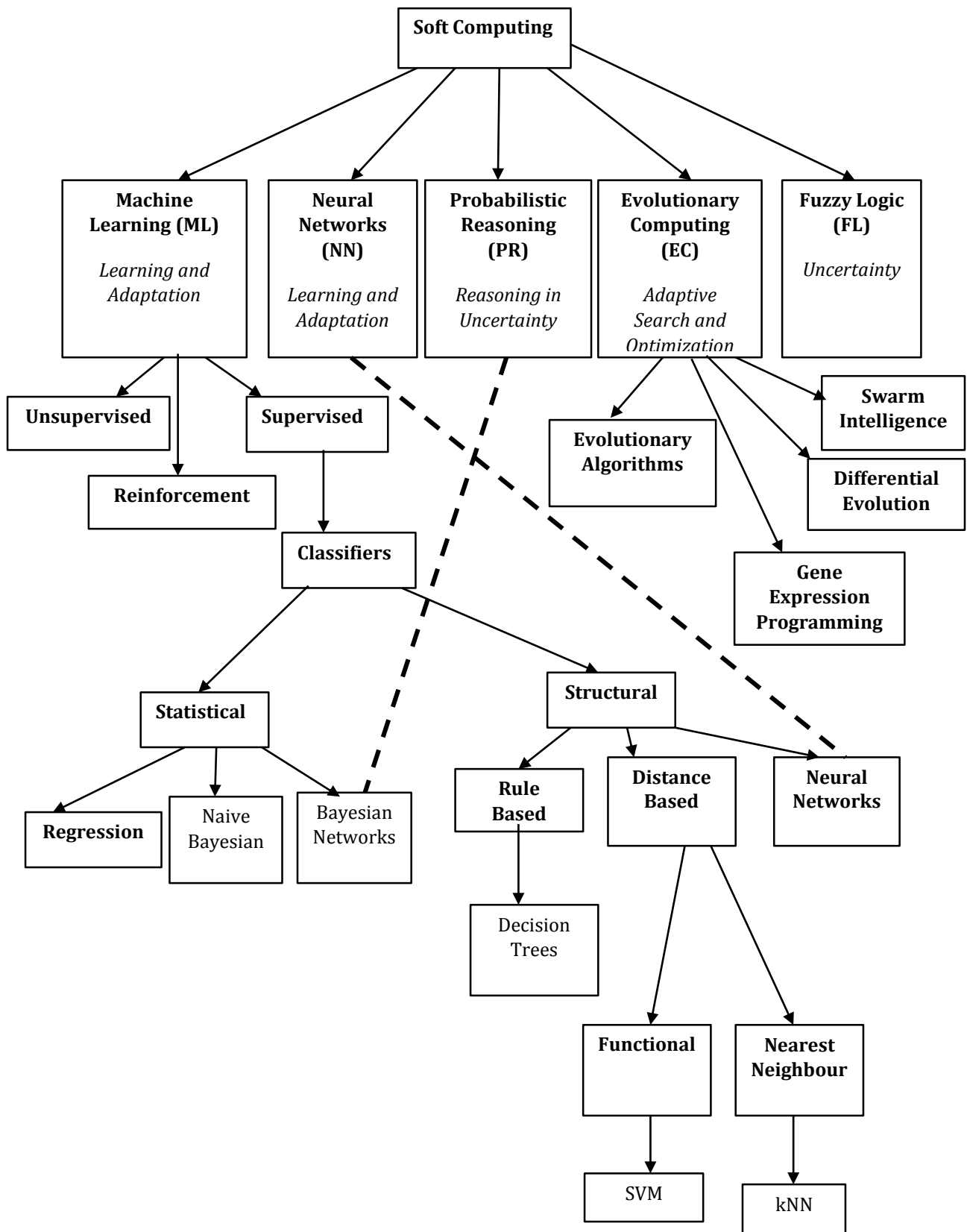
The uncertainty, imprecision and dimensionality in user-generated social media content makes it even more intricate to tap and analyse information using contemporary tools. Novel approaches to information discovery and decision making which use multiple intelligent technologies such as machine learning, deep learning, artificial intelligence, natural language processing and image recognition among others are required to understand data & then generate insights. Thanks to the emerging Soft Computing (SC) techniques [12-14] as they have the ability to capture uncertainty, imprecision and dimensionality in social media content. SC techniques are mainly regarded as optimization techniques that aid modelling real-world problems in order to attain robust & low cost solutions.

Feature selection, noise removal and parameter optimization are few problems which can be solved using soft computing techniques for improved and enhanced sentiment analysis. SC [13] is defined as a collection of computational techniques grounded on artificial intelligence (AI) and natural selection that delivers quick and cost effective solutions to highly complex problems for which analytical formulations does not exists. The term SC was given by Lofti Zadeh in the year 1992. SC targets at finding precise approximation that mainly gives a robust, computationally efficient and cost effective solutions to the problems, eventually saving the computational time as well. Soft Computing (SC) is unique field of study that mainly exploits blending of novel computational techniques that often mimics consciousness & cognition in several vital respects [14]. SC techniques are predominantly considered as optimization techniques that help modelling complex real world problems to achieve robust and low cost solutions. SC has emerged as a significant paradigm to solve real world problems which are pervasively imprecise and uncertain. Application of SC techniques for sentiment classification on social media is a promising direction of research with practical domains for finding, exploring and understanding the extensibility of human expressions.

These techniques are generally divided into the following five categories [12-14]:

- **Machine Learning (ML):** Supervised; unsupervised; or reinforcement learning. Unsupervised includes hierarchical, C means, K means clustering etc.). Supervised includes statistical (regression, naïve bayesian etc.), structural rule based, distance based etc.) and ensemble methods (bagging, boosting, random forest etc.). Deep Learning (DL) is the latest addition to ML and is often regarded as the subset of ML. It's a probable approach used for implementing ML. It includes deep NN (DNN), recursive NN, recurrent NN, convolutional NN, long short term memory and deep belief networks.
- **Neural Networks:** Feed-forward; multi-layer perceptron; artificial neural network, radial basis; kohonen self-organizing; modular; shallow & deep neural networks (DNN), auto-encoder.
- **Evolutionary Computation:** Gene expression programming; differential evolution; evolutionary algorithms (such as genetic algorithms); swarm intelligence (nature-inspired algorithms such as particle swarm optimization, ant colony optimization etc.)
- **Fuzzy Logic:** Classical sets, fuzzy sets, type 1 fuzzy logic, type 2 fuzzy logic, fuzzy arithmetic, membership grade, fuzzy set theoretic operations, fuzzification, de-fuzzification, crisp sets, fuzzy rule base, set theory, if-then else rules .
- **Probabilistic Reasoning:** Bayesian networks (bayesian probability), naïve bayesian, bayes theorem, bayes classifier, multinomial naïve bayes, gaussian naïve bayes, bernoulli naïve bayes.

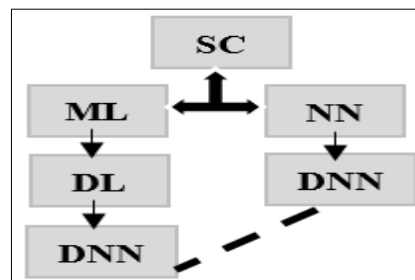
The following figure 1.4 shows the SC techniques and their categorization.



**Fig 1.4.** Types of SC Techniques

The unique property of all SC techniques is their power of self-tuning, that is, they derive the power of generalization from approximating and learning from experimental data [14]. The continuously changing dynamics with respect to increasing user-base and user-activity (posts, comments, likes, re-tweets [19]); trending discussions on topics and issues from varied domains, makes social media; a high-dimensional, complex and fuzzy data space to perform analytics. SC techniques offer a non-trivial solution to the real-world problems which are innately imprecise and uncertain. The guiding principle of SC is to exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness, low solution cost, and better rapport with reality [13]. This generalization for enhanced precision & certainty is normally done in a high-dimensional space and the big social data space of social media presents as the true source for testifying the reasoning and searching capabilities of SC techniques when applied to generic sentiment classification task.

From the observed categorization [13], SC is described as a *'blanket term'* leveraging computational intelligence, comprising of several methodologies which are themselves inter-related to one other in varied forms as shown in figure 1.5.



**Fig 1.5.** Relation between SC, ML and DL

Deep learning (DL) is observed as a sub-part of ML family based on data learning representations. It comprises of techniques such as deep NN (DNN), deep belief networks, recursive NN, convolutional NN and recurrent NN, whereas NN is in itself an established sub-category of SC techniques which consist of feed forward; radial basis; multi-layer perceptron; kohonen self-organizing; modular; shallow and DNN. Therefore, we can infer that SC, ML and DL are inter-related to each other.

SC has appeared as an important paradigm for solving real-world problems which are extensively imprecise & uncertain. Application of SC techniques for SA on social media thus has emerged as an encouraging & promising direction of research with practical domains for finding-exploring-understanding the extensibility of varied human expressions. The alliance of these three domains (social media, SA, SC techniques) suggests the required balancing & necessitates the needed investigation based on feasibility, trends and scope of using SC techniques via *supervised learning* of SA on social media.

## 1.2 Problem Statement & Research Objectives

### **Statement of Research Question:**

*“Can the uncertainty, imprecision and dimensionality of sentiment-rich information on social media be tapped using soft computing paradigms for tractable and optimized results?”*

In response to the identified need for exploiting novel computational techniques for enhanced sentiment analysis on social web, this unifying research question can be broken down into the following three questions, each of which will be addressed by this research:

- How SA on social media facilitates decision support & delivers targeted information?
- What kind of social media data can be mined for SA?
- Which of the SA techniques have been & can be explored to handle and analyse the sentiments in social media?

Consequently, the three main research objectives of the work undertaken are:

- i. **Research Objective I** – To seek the convergence of Web 2.0 technologies and SA on social media for real-life applications.
- ii. **Research Objective II** – To perform SA on textual unstructured data on the Web.
- iii. **Research Objective III** – To propose a novel framework for SA on user-generated content using SC techniques.

## 1.3 Significance of Sentiment Analysis on Social Media using Soft Computing

Proliferation of Web 2.0 has built strong social networks based on user’s personal or professional inclinations. This has eventually intensified the research in the field of text mining specifically SA. *Social media* is often outlined as “VUCCA” (volatile, uncertain, complex, chaotic, & ambiguous) environment which generates vast amount of multimedia content that can be incessantly scrutinized for getting proper insights for market, business and government intelligence leading to superior performances. It has wide range of applications covering almost every domain like social, political, business, management, finance, public, entertainment etc. [21]. Sentiment classification in microblogging portals [21] has become the latest fad for mining opinions for social media analytics. Sentiment classification is often described as a technique of text classification for gauging opinions from huge and diverse user-generated multimedia web-data. Lots of uncertainty is usually associated with the content generated by such microblogs. This is mainly owing to the presence of noisy or heterogeneous data that may be old, outdated, incorrect, ambiguous, incomplete, vague or imprecise etc. Soft computing techniques have demonstrated dexterity in handling uncertainty and imprecision in real-world

problems. *SA or Opinion mining* [23], established as a typical text classification task has always been used to aid knowledge extraction of individual's opinions or sentiments for proficient decision making. It is defined as the computational study of people's opinions, attitudes & emotions towards an entity [5-6]. It is a specialized type of natural language processing (NLP) problem that relies on the analysis of the huge gamut of UGC produced daily via social networks, blogs, e-commerce sites etc. Amongst all social media, Twitter has appeared as a goldmine for testing, reasoning and extending the searching capabilities of mining algorithms when applied to 'generic sentiment classification task' [24]. It has always served as an extensive platform for performing SA with its vast applications studied & reported across pertinent literature [6]. This confluence of sentiment analysis on social media like Twitter etc. has transformed the way computing is done by transforming this "*unstructured UGC*" into a "*sentiment-rich-knowledge-base*" using NLP techniques that can further stimulate effective decision making. One of the key sub-tasks for SA is feature extraction and optimization [25] which when too performed with optimal-parameter setting can lead to enhanced SA. This task is computationally expensive and hard [25]. Furthermore, social media is predominantly an informal way of communication popular among the masses. It comprises of immense usage of slangs, colloquial expressions, mal-formed words, mashed-up words and short forms etc. All this has actually amplified the uncertainty, vagueness & imprecision amongst the available social Web content. Thus, analysing and classifying such enormous, noisy, incomplete and uncertain social web data is often considered as one of the core aspects of opinion mining that indeed affects the quality of sentiments derived and shows-up novel challenges to *feature selection*. As already stated that our studies from the past literature conforms that optimal feature selection leads to enhancement of the classifier's performances, either by performing reduction in the dimensionality, removal of noise and assisting in data visualization for appropriate model selection. Hence in order to boost the predictive sentiment accuracy of such learned models, it is vital to construct an intelligent feature-selection model.

Finding an optimal feature subset is typically intractable and several issues pertaining to feature selection have been shown to be NP -hard [25]. This demands the persistent requisite for examining novel computational methodologies for finding optimal-feature set that could actually improve the performance of the sentiment classifier. One such group of methodologies is called as "SC" which provides a foundation for intelligently mining this huge gamut of "*unstructured user-generated online data*" [13]. Such methodologies bestow robust and low cost solutions for handling and modelling the complexity associated with the real-world problems [15]. Analysing and classifying unstructured Web data that is available online across all social media, using SC methodologies always have a distinctive advantage over machine learning (ML) technologies. It is so because they tend to explore and exploit the human knowledge such as reasoning, recognition and learning into the fields of computing. This follow the possibility of building an intelligent system that majorly are autonomous, self-tuned & automated designed systems. SC, thus, offers an opportunity to signify ambiguity in



human thinking and dealing with the uncertainty in real-life by providing most optimal solution. SC techniques encapsulates the application of machine learning, neural network, evolutionary & swarm intelligence algorithms (SI), fuzzy logic and probabilistic reasoning. Amongst all, machine learning (ML) is being mostly used for sentiment analysis (SA) particularly for classification purposes. SA uses various ML-based techniques for classifying data as positive, negative or neutral depending upon its polarity. Whereas, feature level SA, generally includes the application of many SI-based algorithms for identifying & selecting relevant features from the data.

## **1.4 Organization of Thesis**

This section presents the organization of thesis.

Chapter 1 contains the introduction of fundamental concepts related to research area. The motivation, scope and challenges of the chosen research area has been discussed, followed by a brief description of the key terminologies such as opinion mining, SA, social media and SC. Next it formulates the statement of research problem and decomposes the unified research question, sub questions leading towards certain identified research objectives. A brief description notifying the understanding and significance of SA using SC techniques is illustrated in this chapter. Further, this chapter comprises of organization of thesis as well.

Chapter 2 comprises of a state-of-art systematic literature survey of various application areas of sentiment analysis using SC techniques on social media such as Twitter etc., existing opinion mining techniques and the work done so far in this direction. The work is represented in the form of a *systematic literature review* (SLR) within the promising area of sentiment analysis using SC techniques on social media such as Twitter. Systematic review phases describe the purpose of the stated phases to be performed during the conduct of a SLR. Review planning phase contains the motivation and aim of the research, to gather and analyze the relevant primary studies of research. The next phase i.e. review conduct elaborates searching strategy, literature review of selected studies in visual and tabulated form. Review reporting phase documents the results and discussion of complete review. Thereafter, the identified research gaps are listed followed by the chapter summary.

Chapter 3 explicates about SA on social media using SC techniques in order to seek the convergence of Web 2.0 technologies. The methodologies and findings of first research objective is presented in this chapter, followed by the chapter summary.

Chapter 4 focuses on the application of baseline supervised ML techniques, SI and DL techniques on the user-generated social media content for enhanced sentiment prediction. The methodologies and findings of second research objective is presented in this chapter. The objective of this research is to process the valuable, hidden information from raw, uncertain, imprecise and high-dimensional social media data into a form more amenable to learning and maximizing predictive power using soft computing techniques. Lastly, brief summary of the chapter is presented.

Chapter 5 specifically aims at developing a novel hybrid model for real-time sentiment classification harnessing the best of three diverse domains of soft computing, namely, the deep learning (DL), machine learning (ML) and the swarm intelligence (SI). The methodologies and findings of third research objective is presented in this chapter. A novel model for enhanced sentiment prediction using SI+DL+ML techniques is also discussed in detail. A brief summary of above study ends the chapter.

Chapter 6 recaps the research summary. It also highlights the mapping of the research objectives with list of publications. A thorough discussion of limitations of the study and future scope is discussed. And finally, conclusion at the end winds up the thesis.

Chapters will be followed by a reference lineup which details out the citation sources used in the thesis.

## **1.5 Summary of Chapter**

This chapter has put forward the groundwork for this thesis. It show the ropes of research problem, research objectives and the relevance of using SC techniques as a solution for overcoming the challenges of SA. The need and motivation of the research area has been explained along with the organization of thesis.

# Chapter 2

---

---

## *Systematic Literature Review*

## Chapter 2

# Systematic Literature Review

In order to comprehend a state-of-art within the area of SA using SC techniques on social media, a systematic literature review (SLR) was conducted to review the substantial research.

- This review was planned and conducted based on the format of systematic literature review (SLR) defined by Ketchenham and Charters [26].
- The prime focus was on the understanding the feasibility, scope and trends of SA on social media like Twitter using SC techniques.

Below specified figure 2.1 shows the different phases of SLR. The first phase was referred as formulation of research questions, followed by search strategy and study selection. Next phase was quality assessment and data extraction. Last phase was result reporting.



**Fig. 2.1.** Phases of SLR

The goal of the first phase was to ascertain and formulate the research questions within the domain recognized for survey. Then in the next phase, a search strategy was designed and adopted to ascertain how the search would be conducted. This was primarily done to find and locate the relevant research studies addressing one or more research questions. The scope of the study was narrowed in the study selection phase by using a selection criterion known as inclusion-exclusion criteria. The worthiness of the papers was then calculated using weighted parameters in the quality assessment phase. The purpose of the study selection and quality assessment phase was to ensure the quality and similarity of included studies, and clearly define the boundaries of the review. Post this screening and eligibility decisions on the articles, in the next phase, the data was

extracted to answer the research questions to finally critically analyse the research domain to output a summarized critique which evaluates, extends, or establishes implications for practice, identify gaps and inconsistencies, if any and provide directions for future research.

## 2.1 Formulation of Research Questions

The following sub-sections identify the relevant RQ's which this SLR intends to answer followed by the details of selection and examination of the relevant studies to map studies which address one or more RQ. Following research questions (RQs) were identified to conduct this SLR:

- **RQ1.** On which datasets and domains the studies using SC techniques for SA on Twitter have been conducted?
- **RQ2.** Which are the most frequently used SC techniques for achieving efficient results for the SA on Twitter?
- **RQ3.** What are the widely used performance metrics to evaluate the applied techniques?
- **RQ4.** What is the trend and impact of using SC techniques for SA on Twitter in the past decade?

## 2.2 Search Strategy

A strategy for exhaustive search of all studies that have been meticulously conducted on the topic was set up to find as many potentially relevant papers as possible that relate to the use of SC techniques in SA since the inception of Twitter. For this, first the research questions were broken into individual concepts to create search terms and then databases/ e-portals/ digital libraries to be searched were selected. The search terms were identified like Twitter, sentiment analysis, opinion mining, soft computing techniques, machine learning, supervised methods etc. and were explored in titles, keywords and abstracts of studies to extract all related primary research studies from journals of high reputation and highest relevance to the topic of study, available within five prominent digital libraries (publishers), namely, ACM, IEEE, Elsevier, Wiley and Springer. The grammatical variation of these terms such as synonyms etc. were also used in conjunction with applying wild card for better search or/and boolean expression were used for expanding or narrowing the sweep of the search in order to collect potentially relevant papers. The reference section of the relevant studies was also examined to extract cross-citations. Some secondary studies were also obtained. Thus, the purpose of this step was to identify, select and extract the desired essential subset of research papers for conducting review. This is often called as study selection criteria and process. These extracted studies were then subject to a selection filter which weeded out the irrelevant and redundant papers based on a criterion.

## 2.3 Study Selection

In this phase a selection criteria known as ‘Inclusion-Exclusion criteria’ was adopted to limit the scope of search. It was a kind of relevance filter employed to select or reject studies. The intent was to assess all potential studies which facilitate or directly answer at least one research question within the problem domain. We focus on extracting the research articles based on search terms selected, year of publication, journal specified, citation number of the selected article etc. with the following inclusion-exclusion criteria adopted:

### ***Inclusion criteria:***

- Studies published in journals
- Studies representative of SA specifically to the micro blogging portal Twitter
- Studies focusing on the application of supervised ML algorithms like decision tree (DT), support vector machine (SVM), ensemble methods (EM), k nearest neighbour (kNN), linear regression (LR), logistic regression (LogR), multiple regression (MR) etc.
- Ensemble methods include random forests (RF), bootstrap (BS), stochastic gradient (SGD) etc.
- Studies with supervised learning models in SC such as probabilistic reasoning which includes naïve bayesian (NB), neural networks (NN) (comprising deep NN (DNN), recursive NN, recurrent NN, convolutional NN, long short term memory and deep belief networks), fuzzy logic (FL), evolutionary computing(EC) (containing models like genetic algorithm (GA) etc.) for SA on Twitter.
- Studies with hybrids of SC techniques for SA on Twitter.
- Studies involving the comparison of above mentioned techniques
- Studies involving SA of Twitter in English language only

### ***Exclusion criteria:***

- Studies published in conferences (Though extended versions published in considered journals were included)
- Studies which are without proper empirical analysis or benchmark comparisons
- Studies using any other social media portal like Quora, Facebook, blogs etc.
- Studies with only textual data are considered, other multimedia (image, video and audio) are not included
- Studies that are purely reviews or surveys on SA without any implementations.
- Studies with non-supervised learning model and techniques for implementing SA.
- Studies involving SA on languages other than English and multilingual SA of Twitter (for example languages like Dutch, Portuguese, Latin, Chinese, Arab, Spanish etc.) are not included.

## 2.4 Quality Assessment

In order to maintain the quality standard of the selected studies a careful consideration had been affirmed by taking the novelty of technique proposed and the technical content (data set and evaluation methods used). The quality check had already been ensured as we had only considered selective high quality, high impact journals from reputed digital libraries.

## 2.5 Data Extraction

In this phase finally the key data was extracted from the selected research articles for mapping it to the research questions. The data that was extracted from these studies involved the details about the authors, year of publication, datasets used, techniques applied, domains targeted, type of cross fold validation that was used, key performance indicators that were employed for evaluation of the techniques and the accuracy obtained. All this information was then stored in a table for further data synthesis.

## 2.6 Data Synthesis

The data synthesis phase summarized and interpreted the data extracted to finally output the result of the SLR as direct answers to the identified RQs using critique analysis, discussions and different representations such as tables, graphs, charts, etc. For enhancing the review process, search and study selection procedures were meticulously carried out twice in order to obtain the most relevant and appropriate studies from the literature resources. Research process started by applying the identified search terms on five selected digital libraries that resulted in 502 papers. After removing redundant studies, we obtained 487 studies. Thereafter, application of inclusion and exclusion criteria yielded 60 potentially relevant studies for further analysis. Figure 2.2 depicts the overall search process applied in order to fetch the most relevant studies.

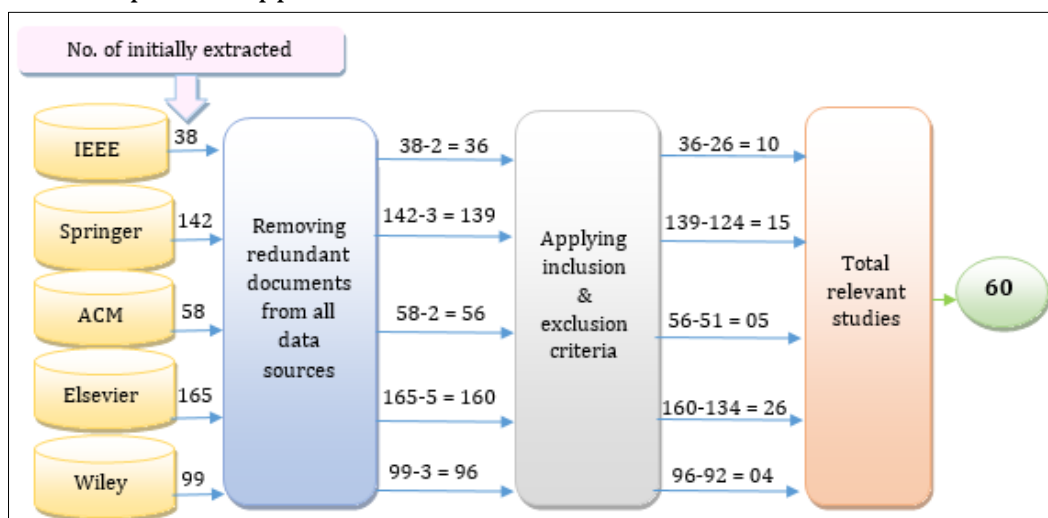


Fig. 2.2. Review Process

The review of literature is presented below in Table 2.1 in a year-wise reverse chronological order. The review of the final set of studies identified for this SLR on use of SC techniques for SA on Twitter is given below. As discussed in the data extraction phase, the data extracted from the selected studies included details about the authors, publication, its year of publication, datasets used, techniques applied, domains targeted, type of cross fold validation used [Ten cross (TC), Five cross (FC), Twenty cross (TWC) etc.], key performance indicators and accuracy obtained.

**Table 2.1:** Summary of the studies undertaken for review

S. No.	Author	Publication	Year	Techniques	Data set	Tools	Domain	Cross validation	Performance parameters	Accuracy
1	Finn et al. [24]	Springer, KI (Künstliche Intelligenz)	2012	NB, SVM, kNN	Author collected 15,000 political and 50,000 non-political tweets in Nov 2012 and in Sept 2012.	Weka	Politics, News	TC	A	NB achieved highest A of 92%.
2	Yerva et al. [27]	Elsevier Journal of Information Systems	2012	NB, SVM	WePS-3	Matlab, openCalais, alchemy API	ACL'08, US census data, different companies.	TC	P, R and F were also used based on the TP, TN, FP and FN (false negatives).	A of 96% of the combined classifiers
3	Lou et al. [28]	ACM Transactions on Knowledge Discovery from Data	2013	SVM	35,746,366 tweets from 10/12/2010 to 12/23/2010.	SVM-light.	Elite users (famous personalities, like actors, singers etc.	TC	P, R and F	TriFG method achieves 27% improvement in comparison to SVM
4	Arias et al. [29]	ACM Transactions on Intelligent Systems and Technology	2013	LR, NN, SVM, DT	Author fetched all the tweets from 20Mar to 20Nov2011, Jun to Aug2011	Weka	Stock market, different companies, movies	-	A, P, R, F, CK	SVM achieved A of approx. 68%.
5	Trilla et al. [30]	IEEE transactions on audio, speech, and language processing	2013	SVM, NB, LogR	Semeval 2007	Weka, EmoLib	News headlines	TC	F	NB shows improvement by 7% for F as compared to the baseline rate.



6	Tuarob et.al [31]	Elsevier Journal of Biomedical Informatics	2014	SVM, RF, NB	700 million tweets from April 2011 to September 2012.	LibSVM, Weka	Health related	TC	P, R, F	Improved P of approx... 63%.
7	Morchid et.al [32]	ACM, Pattern Recognition Letters	2014	SVM, NB	6 million tweets from April 14th 2006 to May 13th of 2011.	Weka	General	TC	P, R, F	SVM produced R of 86.9% and P of 59.8%.
8	Montejo-Ráez et al.[33]	Elsevier, Computer Speech and Language	2014	SVM	376,296 tweets from September 14th 2010 to March 19th 2011.	SVM-Light	General public board messages	TC	P, R, F.	SVM yielded P approx. 64%.
9	Smailovic et.al [34]	Elsevier, Information Sciences	2014	SVM, kNN	1,600,000 tweets prepared by Stanford University. 152,570 tweets discussed about different companies like Apple, Amazon etc. from March 11 to December 9, 2011.	Pegasos SVM	Stock market exchange, companies like Apple, Amazon, Baidu, Cisco, Google, Microsoft, Netflix, people (Bobby Flay, Warren Buffet)	TC	F	The best setting yielded F was approx. 54%.
10	Boella et.al [35]	Springer, Journal of Intelligent Information Systems	2014	SVM	Out of dataset of 1-million Twitter posts, only 100 random were extracted.	Weka	General	TC	P, R, F	P of 66.67 % and a R of 100 %
11	Brynielson et al.[36]	Springer, open journal of Security Informatics	2014	SVM, NB	2.3 million tweet belonging to Sandy hurricane from Oct 29 to Nov.	Weka	Natural crisis related tweets	TC	CM	Improved A of about 60% using SVM.
12	Arakawa et al. [37]	Wiley, Journal of the association for information science and technology.	2014	RF (type of ML that uses DT and creates 1000boots trap(BS) by sampling the variables)	Dataset was created using tweets from 40 accounts of the users who were having high number of followers as of September 7, 2011 containing almost 28,756 tweets	MeCab	categories like 'entertainers (comedians, 'idols,' actors, and sportsmen), 'names' (politicians, authors, cultural icons, intelligentsia	TC	P, R, F	Among all the 15 experiments, experiment 2 showed P of around 84%.

							entrepreneurs, and comic book writers), 'characters,' and 'organizations' (enterprises, local government, and etc.)			
13	Burnap et al. [38]	Springer, Social Network Analysis and Mining (SNAM)	2014	DT, Hybrid: (NB+ LogR=)BLR	Author collected 427,330 tweets over 14 days for the event (i.e. terrorist event 'Woolwich London' 2013),	COSMOS 1 tension engine, SentiStrength tool, VGAM package (R software, Weka	Terrorism	TC	P, R, F	The presence of a URL and a hashtag increased the rate of retweets by a factor of 1.78 (78 %).
14	Makazhanov et al. [39]	Springer, SNAM	2014	NB, LogR, DT	181,972 tweets from 1 Apr 2013 to 11 May 2013 were collected	Weka	Politics (election)	TC	P, R, F	NB and LR achieved P of 93%.
15	Bogdanov et al. [40]	Springer, SNAM	2014	NB	SNAP (dataset contains 467 million posts from Jun to Dec 2009). Author also collected 14.5 million tweets from Mar 2006 to May 2012.	-	Business, celebrities, politics, science and sports.	TC	ER, A	Classifier achieves A of 87 % and lowest ER of 0.13.
16	Lin et al. [41]	Springer, EPJ Data Science	2014	LR, BS	Tweets were collected from 15 Apr to 19 Apr 2013	-	Terrorism	TWC	ER	Error rate of approx. 10% was obtained.
17	Fu et al. [42]	Springer, Neural Computation & Application	2014	FL	Author collected 1,242,522 tweets from 7 Oct 2009 to 13 Nov 2009	-	General	-	Su, Cf	Confidence of approx. 74% was obtained.
18	Chen et al. [43]	IEEE Transactions on Learning Technologies	2014	NB, SVM	Author collected 19,799 tweets and 39,095 tweets from 01 Nov 2011 to 25 Dec 2012 and	LibSVM, Lemur toolkit	Engineering	TC	CK, A, P, R, F, MiAF, MAF	NB outperformed SVM with A of approx. 96% for negative emotions etc.

					05Feb2013 to 17Apr2013.					
19	Liu et al. [44]	IEEE transactions on knowledge and data engineering	2015	SVM, DT, RF	Sanders-Twitter sentiment Corpus. 9413 tweets about 'Taco Bell' during 24-31 Jan2011. 3238 tweets of 2008 Presidential Debate corpus.	Weka	Includes public tweet corpuses about Apple, Google, Microsoft etc.	TC	A, P, R, F	Author's proposed techniques showed improved A of about 82%.
20	Kranjc et al. [45]	Elsevier, Information Processing and Management	2015	SVM	1,600,000(800,000 positive and 800,000 negative)tweets prepared by Stanford University	SVMperf, ClowdFlows, RabbitMQ, Django,LATINO *	General  *(Link Analysis and Text Mining Toolbox)	TC	A, P, R	SVM produces A of 83.01%
21	Sluban et al. [46]	Springer, Computational Social Networks (springer open journal)	2015	SVM	1.6million tweets prepared by the Stanford University. 25,721positive, 23,250negative and 37,951 neutral English tweets. 2,850 positive, 5,569 negative, and 11,439 neutral environmental tweets, from January to December, 2014.	SVMperf, LATINO	General, english, environmental and energy related tweets.	TC	MAF, MAER	For all categories, best model that is the hand-labeled Domain specific model showing highest MAF of 39% and lowest ER of 52.9%
22	Burnap et al. [47]	Wiley, Policy & Internet published by Wiley Periodicals	2015	RFDT, SVM, Hybrid of all above.	450,000 tweets of the event 'murder of Drummer Lee Rigby in Woolwich, London, UK' on May 22, 2013	CrowdFlower, Stanford Lexical Parser, Weka	online hate speech (cyber hate) regarding murder of Drummer Lee Rigby in Woolwich, London, UK in 2013	TC	P, R, F	RFDT (Random Forest Decision Tree) yields R of 55%, SVM yields 69%. P of 89%.
23	Zubiaga et al. [48]	Wiley, Journal of the	2015	SVM	567,452 tweets from 348,757	SVM-light	news, ongoing events,	TC	A, CK	SVM yields A

		association for information science and technology.			varied users for 1,036 unique trending Topics, where tweets being written in 28 different languages like English, Spanish, Portuguese, Dutch, Indonesian etc.		memes, and commemoratives			of around 78%.
24	Magdy et al. [49]	Springer, SNAM	2015	NB, SVM, kNN	Author collected 4, 19.5 million tweets, from end of March to the beginning of May 2014	Weka, SVM Light	Politics, Sports, Entertainment, Science, vehicles	TC	P, R, F, A	SVM has emerge as a best performer among all with 58% P.
25	Tsytsarau et al. [50]	IEEE transactions on knowledge and data engineering	2016	SVM	7 million tweets were collected from 30 trending topics on Twitter from Jun 2009 till Dec 2009.	LK tool, java	General	TC	A, P, R, F	SVM showed A of 78.9 % and authors proposed method depicted the A of 82%. P of SVM is around 91%.
26	Andriotis et al. [51]	IEEE transactions on cybernetics	2016	SVM, NB	Sentiment140 dataset	Weka	Smartphones (Samsung Fame (GT-S6810P)	TC	P, R, F	SVM showed P of approx. 78% for Twitter feeds.
27	Tang et al. [52]	IEEE transactions on knowledge and data engineering	2016	kNN, SVM, NN	Tweets collected were from April 1st, 2013 to April 30 <sup>th</sup> 2013 Urban Dictionary, Twitter dataset from SemEval 2013 and 2014.	LibLINEAR	General	TC	A, MAF	F of SVM is 72.1%for 2013Test and 68.% for2014Test. A of hybrid is 86.1% for2013Test and86.% for2014Test.
28	Peetz et al. [53]	Elsevier, Information Processing & Management	2016	DT	RepLab 2012 and 2013	Weka	Automotive, banking, universities, music	TC	F	RepLab 2013 and 2012 achieved F of 0.55

										and 0.49.
29	Sulis et al. [54]	Elsevier, Knowledge-Based Systems	2016	DT, RF, SVM, NB, LogR	12,532 Tweets (of Task 11) of SemEval-2015.	Weka	Comedians, etc.	TC	F	F of RF for #irony vs. #sarcasm is 69.8%, for #irony vs. #not, is 75.2%, for #sarcasm vs. #not, is 68.4%.
30	Wu et al. [55]	Elsevier, Information Sciences	2016	SVM, NB, LogR	Sanders Twitter sentiment dataset. STS-manual. SemEval 2013.	Matlab R2009b	Companies like Apple, Google, Twitter and Microsoft etc.	TC	A	SVM obtains A of 79%, 82% and 78% for STS, Sanders, and SemEval datasets.
31	Ling Lo et al. [56]	Elsevier, Decision Support Systems	2016	FL, LR, hybrid: SVM+BS, another hybrid:SVM+Bagging.	124,462 Tweets belonged to samsungsg (Twitter account for Samsung Singapore) from 2 Nov 2012 to 3 April 2013. 57,114 tweets of ilovedealssg (Twitter account for daily deals, promotions and discounts in Singapore) from 26 March 2013 to 15 July 2013. 11,969 tweets of beaquafitness (Twitter account of a company focusing on aqua fitness solutions in South East Asia) from 05 Jan 2013 to 11 Nov 2015.	OpenCalais, LibSVM implementation of RapidMiner	Mobiles, fitness, healthy living, daily deals and discounts	TC	P, AP	Hybrid (SVM+EM) obtained F of 98% for samsungsg. 97% for ilovedealssg dataset, 98% for beaquafitness dataset.

32	Zoonen et al. [57]	Elsevier, Computers in Human Behavior	2016	SVM, NB, LogR	578,803 tweets in dutch language being sent by 443 employees who worked in various organizations and work with an average of 39.62 hours per week for an organization having at least thirty employee.	scikit-learn	government /public administration, education/science, health care, business services, trade/commercial services, industry, financial services etc.	TC	A, R, P, AUC	SVM achieved A of 81%.
33	Wang et al. [58]	Springer, Eurasip Journal on Wireless Communications and Networking	2016	SVM, kNN, LogR, NB	Author collected tweets from Twitter between Mar. 1st and May 1st of 2015	scikit-learn	News related subjects	FC	P, R	SVM has achieved A and R of 89.8% and 89%.
34	Celli et al. [59]	Elsevier, Information Processing and Management	2016	LogR, RF	Gold Standard,	-	News	TC	P, R, F	61.7 % F was obtained.
35	Igawa et al. [60]	Elsevier, Information Sciences	2016	RF, NN	Author collected all the tweets of the FIFA World Cup 2014	-	Sports	TC	A, P	RF achieved 88.7% A.
36	Korkmaz et al. [61]	Springer, SNAM	2016	LogR	500 million tweets were collected from Nov 2012 to Aug 2014.	-	Social and political	-	P, R, F	Average F scores are in range 0.68-0.95.
37	Burnap et al. [62]	Springer, EPJ Data Science	2016	SVM, RF	1803 tweets of sexual orientation, 1876 tweets of racism, 1914 tweets of disability were collected on 30Apr2013, 6Nov2012, 29Aug2012.		Cyber hate	TC	P, R, F	The mean precision of the individual classifiers for cyber hate was 0.85.
38	Oliveira et al. [63]	Elsevier, Expert Systems With Applications	2016	MR, NN, SVM, RF	Author collected 31 million tweets from 22Dec2012 to 29Oct 2015.	R-Tool, Mongoddb, Stanford CoreNLP	Stock Market	TC	NMAE	SVM majorly produced the most accurate results.

39	Perikos et al. [64]	Elsevier, Engineering Applications of Artificial Intelligence	2016	NB	Author collected 250 random tweets	Python's NLTK toolkit	News	-	P, A, Sp, Sn	NB achieved A of approx. 85%
40	Brocardo et al. [65]	Wiley, International Journal for communication system	2016	SVM	Author collected 3194 tweet as on before 6Nov2013	Weka	General	TC	FAR, FRR, EER	The best setting yielded error rate as 10.08%
41	Bouazizi et al. [66]	IEEE Access	2016	SVM, kNN, RF	Author collected 7628 tweets from Dec2014 to Mar2015.	Apache OpenNLP, Weka, LibSVM	General	-	A, P, R, F	SVM obtained the highest P of 98%.
42	Farias et al. [67]	ACM Transactions on Internet Technology	2016	NB, DT, SVM	Author collected more than 30,000 sarcastic or ironic tweets	Weka	Education, humor, politics, news	TC	F	SVM achieved highest F of 0.90
43	Sintsova et al. [68]	ACM Transactions on Intelligent Systems and Technology	2016	NB, LogR	Author extracted 52218 tweets	Weka, LibLINEAR	Sports	-	A, P, R, F, MAF	Highest macro R is obtained by LogR
44	Nair et al. [69]	Elsevier, Computers and Electrical Engineering	2017	DT	processed.cleveland.data from Heart Disease Data Set of UCI machine learning repository where a user tweets the necessary data like age, heart rate etc.	Apache Spark's machine learning library, MLlib written with Scala	Health (heart diseases)	70:30	A (Author took 70% data for training and 30% for testing.)	It yields higher accuracy in less time and free of cost.
45	Cui et al. [70]	Springer, Expert Systems With Applications	2017	SVM	132.6 million tweets in April, May and June 2014 by 23.2 million accounts were considered	LibSVM	NGOs, charities, events, journalists/bloggers (freelance media professionals or news agencies), celebrities, politicians, sportsmen etc.	80:20	P, R, A, F (Author took 80% data for training and 20% for testing.)	Improved results were obtained using enhance distant based supervised algorithm along with SVM.
46	Gállego et al. [71]	Elsevier, Information Fusion	2017	NB, LogR, SVM	1,600,000 tweets with emoticons were collected from April 6,	LibLINEAR and LibSVM	General	FC	GM	Improved results were obtained using NB.

					2009 to June 16, 2009.					
47	Alsine et. al [72]	Elsevier, International Journal of Approximate Reasoning	2017	SVM	Author collected tweets belonging to the month of Mar 2016 and Apr 2017	natural language Toolkit (NLTK)	Taxation, politics, public campaign	-	A	Improved A approx. 60% obtained for SVM.
48	Jianqiang et al. [73]	IEEE Access	2017	SVM, NB, LogR, RF	Stanford Twitter Sentiment, SemEval2014, Stanford Twitter Sentiment Gold (STS-Gold), Sentiment Strength Twitter (SS-Twitter), Sentiment Evaluation (SE-Twitter)	scikit-learn	-	TC	A, F	NB achieved highest F of 0.37 for SemEval 2014.
49	Jain and Kumar [74]	Journal of Computational Science, Elsevier	2017	SVM, NB, LogR	Author collected tweets from Sept 2016 to Nov 2016 using Twitter API.	LibLinear	Health	TC	F, P, R, A	Performance of SVM was observed to be better as compared to NB.
50	Keshavarz and Abadeh [75]	Knowledge-Based Systems, Elsevier	2017	GA	Author had utilized benchmark datasets, namely Sanders, Presidential debate corpus, Healthcare Reform (HCR), SemEval 2013 and Stanford.	-	Companies like Apple, Google, Twitter and Microsoft etc., Politics (Elections), Health.	TC	F, A, P, R	Accuracy of more than 85% was obtained.
51	Xiong et al. [76]	Neurocomputing, Elsevier	2017	SGD, SVM, NB-SVM (NB enhanced SVM), NN, CNN	SemEval 2013			TC	MAF	CNN yielded improved results with MAF score of around 85%.
52	Neppalli et al. [77]	International Journal of Disaster Risk Reduction, Elsevier	2017	NB, SVM	Author collected geo-tagged tweets from Hurricane Sandy collection comprising	SentiStrength	Environmental crisis	TC	A	SVM produced enhanced results with A of around 76%.



					of 74,708 tweets with geo-location.					
53	Singh et al. [78]	Transportation Research, Elsevier	2017	SVM, NB	Author collected random 10,664 tweets using Twitter handlers.	-	Food (health)	FC	A	It was observed that performance of SVM was better than NB.
54	Xiaomei et al. [79]	Knowledge-Based Systems, Elsevier	2017	SVM, NB,	Author had utilized benchmark datasets, namely Sanders, Presidential debate corpus and Healthcare Reform (HCR).	-	Health, Obama, Republicans, Democrats, conservatives, liberals, elections, politics, Tea Party	FC	A	SVM yielded improved accuracy.
55	Khan et al. [80]	International Journal of Information Technology, Springer	2017	NB	Author collected 20,000 political and non-political.	Ipython notebook, Apache Spark, Python nltk library	Politics	-	P	Author proposed algorithm achieved highest A of 85 %.
56	Bouazizi and Ohtsuki [81]	IEEE Access	2017	RF	Author collected 21,000 tweets for training and 19740 tweets for testing purposes.	SENTA	Random	FC	A,P,R, F	RF achieved A of 60.2% for multi-class SA.
57	Li et al. [82]	Information Systems, Elsevier	2017	NB, DT	Author collected 196,370 tweets and classified them into 20 classes.	MongoDB, Java	Stock Market	TC	A	NB yielded A of more than 72%.
58	Jianqiang et al. [83]	IEEE Access	2018	SVM, CNN	Author used benchmark datasets namely STS-Test, STS-Gold, SS-Twitter, SE-Twitter.	GloVe	Random	TC	A, P, R, F	GloVe-CNN achieved highest A of around 87.62% using STS dataset.
59	Ghiassi and Lee [84]	Expert Systems with Applications, Elsevier	2018	NN, SVM	Author collected around 40,000 tweets from 8th Jan 2013 to 11th April 2013 related to Starbucks, Governor Christie, Southwest	WEKA, Java, MS SQL Server	Consumer products and services, Politics, Entertainment.	DAN2	P, R, F	Author achieved domain transferability for different datasets. SVM yielded enhanced results.

					airlines and Verizon.					
60	Symeonidis et al. [85]	Expert Systems with Applications, Elsevier	2018	NB, SVM, LogR, CNN	Author used benchmark datasets namely SS-Twitter and SemEval 2013-2017.	NLTK, Sklearn	Random	-	A	Author achieved best results with CNN.

## 2.7 Literature Survey

In 2012, Finn et al. [24] proposed the use of ML techniques like NB, SVM, kNN for labelling the tweets under 'political activist' and 'general public' categories. NB outperformed the other two with an accuracy of around 92%. Another study was proposed by Yerva et al. [27] in 2012 who discussed about the general entity matching issue pertaining to Twitter message classification using hybrid classifiers like NB and SVM for Twitter using WePS-2, 3 dataset on the domain like ACL'08 (Association for Computational Linguistics Program committee members), US census data, different companies. They had combined classifiers at the pre-processing level and had observed that the combined classifiers obtained improved A of 96%.

In 2013, Lou et al. [28] proposed a novel method to formulate triadic closure social relationships on Twitter and had compared it with classifiers like SVM etc. they had observed that their approach produces A of 90% for reciprocal relationships on Twitter in comparison to SVM etc. Another author Arias et al. [29] (2013) focused on the applicability of LR, NN, SVM, DT techniques for building and evaluating the forecasting models for different time series data sets under different experimental conditions for Twitter. Trilla et al. [30] (2013) worked majorly on the implementation of the unigrams so as to adapt to the SA methods with successful classifiers like SVM, NB, LogR for the news headline domain of Twitter using Weka and had observed that these methods yield better and improved results in comparison to bigrams etc. when used with successful classifiers like SVM.

In 2014, Tuarob et.al [31] had discussed the usage of techniques like SVM, RF, NB for health related tweets using varied feature sets. SVM is found to perform the best with improved and better F measure of 68.47%. Morchid et.al [32] (2014) worked towards the analysis and detection of massively retweeted tweets on selected features using techniques like SVM, NB. Improved results were achieved when SVM. A novel approach was presented by Montejó-Ráez et al. [33] in 2014, for calculating the scoring of tweets according to their polarity using random walk analysis and comparing it with SVM. Authors approach produces P of approx. 63% in response to SVM which yielded P of approx. 64%. Author Smailovic et.al [34] (2014) proposed an incremental active learning approach for increasing the predictive power of sentiment classifiers for stock market

with improved results via applying NB, SVM, kNN. It was observed that NB had lower performance in comparison to SVM. Boella et.al [35] (2014) proposed an approach for extraction of semantic information automatically from tweets. Then such extracted tweets are fed in to SVM for providing 'semantic-aware search queries'. Author achieved P of 66.67 % and R of 100 % for the definitional tweets. In 2014, a methodology was developed by Brynielsson et al. [36] for selecting and collecting natural calamity related tweets and tagging them using the annotators like happiness, anger, fear etc. They had focused on the applicability of the classifiers like SVM, and had observed that SVM outperformed NB with an improved A of about 60%. In 2014, the author Arakawa et al. [37] had proposed tweet classification random forests experimentation that analyses number of retweets and the effects of the features based on the 'user type'. In their work, the results claimed that the 'classification by user type' depicted the overall best performance. It was also observed that information roles and function words are important aspects in the retweeted classes and analysing number of features is important in determining user types. Among all the 15 experiments, experiment 2 showed precision of around 84%.

A model was built by Burnap et al. [38] (2014) that predicted information flow size and survival on Twitter following a terrorist event via the action of retweeting using DT, BLR techniques. The novel findings were the time lags between retweets, the co-occurrence of URLs and hashtags, and the sentiment expressed in the tweet. Makazhanov et al. [39] (2014) predicted user political preference from their Twitter behaviour towards political parties for 2012 Albertan and 2013 Pakistani elections. They build prediction models based on a variety of behavioural and contextual features using NB, LR, DT techniques. A genetically inspired framework was proposed by Bogdanov et al. [40] (2014) for modelling individual social media users which they termed a genotype. They extracted topic-specific influence backbone structures based on content adoption and further showed that genotype model with combination of NB enable more than 20 % improvement. Lin et al. [41] (2014) studied the expression of fear and social support in Twitter communication during and after a terrorist attack using methods like LR, BS. Using nearly all geo-tagged tweets they had examined the temporal correlation in these expressions. Their findings suggests that not all fear is necessarily bad and could be considered interesting for general prospects of terrorism as a strategy for political change in the era of social media. A novel method was proposed by Fu et al. [42] (2014) for extracting useful behavioral trends of users on Twitter using the mass assignment theory based fuzzy association rules. The paper uses FL in developing the new scheme and gave improved results. Another work focusing on the usage of techniques like NB, SVM etc. was proposed by Chen et al. [43] (2014) for depicting the student understandings, issues, problems and challenges faced by them during their studies using social media like Twitter. In 2015, a model was developed by the author Liu et al. [44] for fetching unlabeled tweets from a mixed group of labeled and un-labeled tweets for maintaining the dynamism of the selected tweets and classifying them depending on the trends of the topics selected with improved A using SVM, DT, RF classifiers. Kranjc et al. [45] (2015)

focused on the implementation of the active learning scenario' for Twitter using cloud based data mining platform with improved performance by classifying tweets as positive and negative only i.e. two way classification. SVM produces A of 83.01%. Sluban et al. [46] (2015) had proposed a work that divided sentiment model into three categories as Smiley-labeled general, Hand-labeled general, and Hand-labeled domain-specific sentiment model utilizing the already processed negative and positive tweets. They had further proved that the 'high-quality domain specific tweets' provides a much better sentiment model despite the number of available tweets for it. They had implemented SVM. Author observed that among all the three categories stated, best model that outraged others is the hand-labeled domain specific model that showed lowest ER of 52.9% and the highest MAF of 39% on the test set. A classifier was developed by Burnap et al. [47] (2015) for monitoring public reactions to emotional hateful events like death of Rigby using SVM, RF, DT and hybrid of these. Prime aspect of the author was to evaluate the hybrid classifier which could be further used by policymakers for effective and efficient decision making process for such cyber hate on social media. Based on the implementation results, author lead to the conclusion that ensemble classification process is most effective and efficient for classification of such cyber hate events, provided the current feature sets. Zubiaga et al. [48] (2015) proposed a method so as to efficiently categorize trending topics irrespective of the need of any external data using SVM and it was observed that SVM yields A of around 78%. Magdy et al. [49] (2015) experimentally demonstrated the effectiveness of a "distant supervision" approach to tweet classification, consisting in automatically obtaining labelled data from one social media platform (YouTube) and using this data for training a classifier for another such platform (Twitter) using kNN, SVM, DT techniques.

Tsytsarau et al. [50] (2016) had focused on the aggregation of the large real-time datasets having diversified sentiments and thereafter had developed a model that performs the sentiment contradiction diversification at different time scales. Author had used a novel data structure which is incrementally maintained and helps in scaling large amount of datasets, often called as contradiction tree. The results claim that the SVM had shown improved results in terms of measuring the contradiction level and ranking of the dataset. In 2016, another study depicting the investigation of the effect of textual data on the short message services (SMS) so as to perform SA on the smart-phone users for revealing the mood trends in them and comparing them with the Twitter feeds was given by Andriotis et al. [51]. They had majorly focused on the data being stored in the internal storage of our smart phones and illustrating inter-connections within the entities at all the levels of the ecosystem and their approach primarily targets the data that is found in the smart phones, which is linking the users to the universal digital community. The results claimed that the P and F score of SVM was just more or less comparable to other classifiers being implemented by the author like NB. Tang et al. [52] (2016) had implemented a recursive neural network and convolution neural network with dedicated loss functions so as to record sentiments of sentences or words as well as contexts of those words for learning word embedding i.e. author had developed NN based ranking

model for learning sentiment embedding by utilizing sentence level sentiment information as 'task-specific evidences'. In their work, author had used urban dictionary in order to make clusters of all related words together and then applied kNN classification to these clusters so as to classify them into positive, negative and neutral clusters of high quality similar words sentiments. Peetz et al. [53] (2016) had focused on the estimation of the polarity of the tweets in terms of reputation, for which DT was used for combining, learning and finding the optimal number of features in a set. In the preliminary experimentation, author had observed that the SVM and RF showed poor performances in comparison to DT for varied selected features. The results exclaimed that the DT was more successful in making decisions when applied to the tweets belonging to domains like automotive, banking, universities, and music etc.

Sulis et al. [54] (2016) had briefly discussed about the figurative content of the tweets such as hashtag's 'for not, sarcasm and irony' by applying techniques like DT, RF, SVM, NB, LogR. Author's aim was to explore all the differentiating traits among these figurative tweet contents. In their work, it was observed that the RF gained the highest F score and DT got the lowest F score for all the #tag combinations for figurative intent tweet messages. The best result had been observed by RF for the case of #irony vs. #not classification, being approx. 75.2% which profoundly provides better insight into the use of these types of hashtags for labelling the tweets (whether they are ironical, or sarcastic etc.) being Twittered on social media. In 2016, author Wu et al. [55] had worked towards extraction of the useful sentiment oriented knowledge from the unlabelled tweets in order to improve and enhance the microblog sentiment classification using SVM, NB, LogR. The experimental results proved that the author's approach improved the process of sentiment classification effectively by reducing the dependency on the labelled data. Ling Lo et al. [56] (2016) had worked towards establishment of the top-n followers and ranking them which could eventually help the companies (belonging to mobiles, fitness, healthy living, daily deals and discounts) and promoters to publicize their businesses on Twitter. Author had implemented the concept by applying methodologies like LR, Fuzzy logic, hybrid: BS ensemble using SVM model, another hybrid: bagging ensemble using SVM models. Zoonen et al. [57] (2016) proposed an approach that enabled to perform the analysis of the entire tweet texts and thereby helped in reducing the risk involved with the sampling errors. Author established that it could be applied to other social media content as well for varied topics in demand using multiple method approaches. SVM outperformed the methods like NB, LogR and produced acceptable and higher level values for all the performance parameters when applied to it thus yielding highest reliability statistics. In 2016, another author sisWang et al. [58] had focused on the determination of 'multivariate emotional model classification'. Author had also applied deep learning for the 'entity recognition' via using 'SENNA' deep learning toolkit. Author applied classifiers like SVM, kNN, LogR, NB. SVM has emerged to give most promising results in comparison to other methods applied. It produced a lower error rate and higher accuracy. SVM has achieved accuracy and recall of 89.8% and 89% respectively which is higher in regard to other classifiers.

Celli et al. [59] (2016) had analysed the role of personality and communication styles in the diffusion of news articles using LR, RF. They had automatically annotated personality types and communication styles of Twitter users and analysed the correlations between personality, communication style, Twitter metadata (such as following and followers) and the type of mood associated to the articles they shared. Another study demonstrating a wavelet-based approach for account classification was given by Igawa et al. [60] (2016) that detects textual dissemination by bots on an Online Social Network. Their main objective was to match account patterns with humans, cyborgs or robots, improving the existing algorithms that automatically detect frauds. Experiments were performed using a set of posts crawled during the 2014 FIFA World Cup, obtaining accuracies within the range from 94 to 100% via RF and NN (multilayer perceptron). Korkmaz et al. [61] (2016) had presented a model for predicting civil unrest through the combination of heterogeneous online data sources and provide a critical evaluation of the approach via implementing LogR. They had evaluated the predictive power of disparate datasets and methods, and provide interpretable insights into unrest events. Another author Burnap et al. [62] (2016) had developed novel machine classification models to identify different types of cyber hate individually. The resulting cyber hate classification models have been shown to be applicable to a range of protected characteristics including race, disability and sexual orientation, and provide new ability to automatically identify content perceived by a group of human annotators as hateful or antagonistic. They had implemented SVM and RF. A model was proposed by Oliveira et al. [63] (2016) for assessing the impact of tweets on stock market variables such as returns, volatility, etc. The author had applied ML techniques such as MR, NN, SVM and RF to detect whether the predictions based on sentiments are influential on the stock market or not. Another study for deriving dependencies and the emotional state of the sentences was put forward by the author Perikos et al. [64] (2016). This ensemble classifier was implemented on a dataset of news headlines and Twitter posts to reveal the best performer with higher A and P using NB. A stylometric analysis technique in continuous authentication was explored by Brocardo et al. [65] (2016). It proceeds by breaking an online document into a sequence of short texts on which the CA decisions happen. The method yielded promising results with an equal ER varying from 8.21% to 16.73%. Bouazizi et al. [66] (2016) proposed the implementation of techniques like SVM, kNN, RF for detection of sarcastic comments using pattern based features. Farias et al. [67] (2016) proposed the usage of techniques like NB, DT, SVM for differentiating between the ironic and the non-ironic content using affective information. Sintsova et al. [68] (2016) had focused on the usage of NB and LogR for building emotion classifiers.

Nair et al. [69] (2017) had developed a health monitoring application for prediction of heart diseases based on spark cluster ML model using DT methodology for prediction of health status of an individual by harnessing real-time data from Twitter. The application has been deployed using Cloud in 'Amazon Elastic Compute Cloud (EC2)'. Cui et al. [70] (2017) had formulated the use of SVM together with other distant

supervised classification algorithms for classifying Twitter accounts as Branding and Personal account types without the involvement of any manual labelling. Gállego et al. [71] (2017) had focused on the usage of the EM together with NB and SVM for incorporating binary quantification. The results showed that the better performances were obtained by using the ensemble versions. Alsine et al. [72] (2017) had presented a model based on valued abstract argumentation for automatically labelling the relationship between the sentiments via implementing SVM. It also reasons about the accepted and the rejected sentiment tweets for the controversial discussions on Twitter. Jianqiang et al. [73] (2017) discussed about the performance of the classifiers like NB, SVM, LogR, RF on five different benchmark Twitter datasets. The results indicate that NB and RF are more sensitive in comparison to the other classifiers for varied pre-processing methods. Jain and Kumar [74] in 2017, had applied SVM, NB and LogR to health domain for SA. The results were evaluated using precision, accuracy, recall and F measure parameters. The highest accuracy was obtained by SVM. Keshavarz and Abadeh [75] (2017) demonstrated the applicability of genetic algorithm (GA) model for SA to benchmark datasets namely, Sanders, Presidential debate corpus, Healthcare Reform (HCR), SemEval 2013 and Stanford. Improved results were obtained using GA. Xiong et al. [76] (2017) applied soft computing techniques such as stochastic gradient descent (SGD), SVM, NB-SVM (NB enhanced SVM), MLP and CNN to SemEval 2013 benchmark corpus using ten-fold cross validation. CNN yielded improved results with MAF score of around 85%. Neppalli et al. [77] (2017) collected geo-tagged tweets from Hurricane Sandy Collection for analysing sentiments of the tweets belonging to the environmental crisis. Author gathered around 74,708 tweets with geo-location using SentiStrength and applied SVM and NB. SVM produced enhanced results with A of around 76%. Another similar work was given by Singh et al. [78] in 2017 where the author applied SVM and NB to analyse sentiment polarity of the tweets belonging to health domain. It was again observed that performance of SVM was better than NB. Xiaomei et al. [79] (2017) utilized benchmark datasets, namely Sanders, Presidential debate corpus and Healthcare Reform (HCR) for SA using five-fold cross validation. The tweets belonged to varied domains such as Health, Obama, Republicans, Democrats, conservatives, liberals, elections, politics etc. SVM yielded improved accuracy. Khan et al. [80] demonstrated the applicability of NB for SA by collecting around 20,000 political and non-political tweets as dataset. The results were evaluated using precision efficacy criterion. Bouazizi and Ohtsuki [81] (2017) applied ensemble method namely RF for SA. Author collected 21,000 tweets for training and 19740 tweets for testing purposes using FC validation technique. RF achieved A of 60.2% for multi-class SA. Li et al. [82] (2017) focused on the use of NB and DT for SA for the tweets belonging to stock market exchange domain. NB yielded A of more than 72%.

Jianqiang et al. [83] in 2018 applied SVM and presented a deep neural network model namely CNN for SA using GloVe. Author utilized benchmark datasets namely STS-Test, STS-Gold, SS-Twitter, SE-Twitter for SA. GloVe-CNN achieved highest A of around 87.62% using STS data set. Ghiassi and Lee [84] (2018) analysed sentiment polarity using techniques such as NN and SVM. Author collected around 40,000 tweets from 8th Jan

2013 to 11th April 2013 related to Starbucks, Governor Christie, Southwest airlines and Verizon. Author achieved domain transferability for different datasets. SVM yielded enhanced results. Symeonidis et al. [85] (2018) applied soft computing techniques namely NB, SVM, LogR and CNN for analysing sentiments. Author used benchmark datasets namely SS-Twitter and SemEval 2013-2017. Amongst all, CNN yielded best accuracy.

Thus, we can infer that very few studies exist that demonstrate the application of swarm and DL methods for SA on Twitter, making it completely a potential area for research and development in the field of SA. Researchers and academicians are open to substantiate the influence of swarm and DL methods for sentiment analysis on social media such as Twitter.

## 2.8 Key Observations and Research Gaps

The SLR enabled uncovering some common observations and important trends in the research area. The following **research gaps** were identified:

- Analysing and classifying sentiments from web data in natural language is challenging as effective feature selection is difficult & computationally expensive.
- Many problems related to feature selection are NP -hard and finding an optimal subset of features is usually intractable. For datasets which are smaller in size, it is still manageable but for the large datasets, manual feature extraction is quite tough. This necessitates examining new computational methodologies for finding optimal feature set which performs the automatic extensive feature extraction, improves the performance of the sentiment classifier in terms of predictive accuracy and result comprehensibility.
- For datasets which are smaller in size, it is still manageable but for the large datasets, manual feature extraction is quite tough.
- This necessitates examining new computational methodologies for finding optimal feature set which performs the automatic extensive feature extraction, improves the performance of the sentiment classifier in terms of predictive accuracy and result comprehensibility.
- Though researchers are keen in applying soft computing techniques, only few approaches have been explored. Techniques such as deep learning, evolutionary computing, optimization algorithms and hybrid approaches including neuro-fuzzy models have been least explored or implemented to substantiate their influence on sentiment analysis.
- Analysing sentiment analysis using fuzzy logic approaches like type 2 fuzzy models is yet another novel dimension that is open for further exploration.
- Also, nature inspired algorithms (NIA) including swarm intelligence algorithms and bio-inspired algorithms like flower pollination, grey wolf, moth flame etc. have not been



implemented yet in order to signify their impact on sentiment analysis when examined for varied social media.

- Fine-grain sentiment analysis which includes, emotion analysis, sarcasm detection, rumour detection, irony detection have been identified as potential directions of research.
- Social media has become an informal way of communication with accelerated use of slangs and emoticons, mal-formed words, colloquial expressions and multilingual content thus increasing the dimensionality, fuzziness and complexity of the content.
- At the same time, most real-time datasets scrapped from social media suffer from imbalance class distribution (skewed data), that is, the dataset is imbalanced. This amplifies the uncertainty and imprecision within the available social web content.
- To increase the predictive sentiment accuracy of learned models, it is imperative to build an intelligent feature selection model that could accurately and efficiently classify sentiments.
- Incessant need to enhance the performance of the sentiment classification tools which are now in practical use within various business and social domains.
- The tools and software are useable and affordable only by organizations (both private and government) but currently unavailable to generic users for assisting intelligent and personalized data analysis.

The existing research gaps in SLR urges the need to exploit new computational techniques for improving the sentiment classification accuracy on social web is identified making this domain of study a potentially active and dynamic for both researchers and practitioners. Hence, we can infer that a multi-prolonged approach utilizing the self-tuning capabilities of SC techniques is required that could aid in enhanced prediction of sentiment polarity.

## **2.7 Chapter Summary**

This chapter presented a systematic and comprehensive literature review on the research work done in different application areas of SA on Twitter using SC techniques. Some important conclusions have been drawn by answering identified research questions. The SLR helped us to identify the research gaps within the selected domain and aided in giving us various research directions to work upon.

# Chapter 3

---

## *Sentiment Analysis using Soft Computing: Convergence with Web 2.0*

## Chapter 3

# Sentiment Analysis using Soft Computing Techniques: Convergence with Web 2.0

Sentiment detection and classification is the latest fad for social analytics on Web. With the array of practical applications in healthcare, finance, media, consumer markets and government, distilling the voice of public to gain insight to target information and reviews is non-trivial. With a marked increase in the size, subjectivity and diversity of social web-data, the vagueness, uncertainty and imprecision within the information has increased manifold which further makes it crucial to tap and analyse information using contemporary tools. Text mining models define the process to transform and substitute this unstructured data into a structured one for knowledge discovery.

Studies from our pertinent literature and practical applications in the field of SA have escalated in the past decade with the transformation and expansion of Web from passive provider of content to an active socially-aware distributor of collective intelligence. This new collaborative Web (called Web 2.0) [6], extended by Web 2.0 technologies (Web-based technologies) like comments, blogs and wikis, social media portals like Twitter or Facebook, that allow to build social networks based on professional relationship, interests, etc. encourages a wider range of expressive capability, facilitates more collaborative ways of working, enables community creation, dialogue and knowledge sharing and creates a setting for learners to attract authentic audiences by various tools and technologies. It has expanded the scope of commercial activities by enabling the users to discuss, share, analyse, criticize, compare, appreciate and research about products, brands, services through social platforms like Voonik, Facebook, and Twitter etc. This pool of information can be explored for a mutual benefit of both the customer and the organization. Data Analytics on these social web-based corpora has thus been an ongoing trend where online comments are transformed into a sentiment-rich knowledge-base that can leverage efficient and effective decision making.

Amongst the Web 2.0 tools, Twitter has evolved as a major revolution in field of social media and has a global reach. It has been the most preferred social channel from which sentiment rich data can be extracted. Moreover, the increased dimensionality, complexity and fuzziness in the user-generated Twitter data further fosters the need to look for improved and optimized sentiment classification techniques. Studies are constantly being conducted to explore new SC paradigms which handle uncertainty, imprecision, approximation, partial truth, fuzziness and allow replication of human intelligence for personalized and tractable results.

Therefore we seek 3.1. The methodologies and findings for the first research objective is presented in this chapter. A brief summary of above study will ends the chapter.

### 3.1 Research Objective 1

**Research Objective:** To seek the convergence of Web 2.0 technologies and sentiment analysis on social media for real-life applications.

### 3.2 Methodology

This research primarily aims to study SA on various social media for data-driven decision making on real-life applications. Sentiment detection and classification is the latest fad for social analytics on Web. SA has emerged as one of the most dynamic area of research in recent times. SA makes it viable to instantly transform unstructured data from social media sources like Twitter into structured data to comprehend intelligence. With a marked increase in the size, subjectivity and diversity of social web-data, the vagueness, uncertainty and imprecision within the information has increased manifold. Soft computing techniques have been used to handle this fuzziness in practical applications. SC techniques offer a non-trivial solution to the real- world problems which are innately imprecise and uncertain. The unique property of all SC techniques is their power of self-tuning, that is, they derive the power of generalization from approximating and learning from experimental data [13]. This generalization for improved precision and certainty is usually done in a high-dimensional space and the big social data space such as Twitter serves as the true source to testify the reasoning and search capabilities of SC techniques when extended to a generic sentiment classification task. Continuously changing dynamics with respect to increasing user-base and user-activity (posts, comments, likes, re-tweets); trending discussions on topics and issues from varied domains, makes Twitter a high-dimensional, complex and fuzzy data space to perform analytics.

To study, explore and analyse the existing work on SA on social media specifically on Twitter using soft computing techniques and to report gaps, future directions in the research area, a thorough review of literature was conducted. The important practices followed for conducting such a review include surveys, narrative reviews, systematic literature review (SLR) and meta-analysis. The SLR was chosen for review in this research. An SLR seeks to systematically search for, appraise and synthesize research evidence often adhering to the guidelines on the conduct of review. Format of SLR given by Ketchenham and Charters [26] was adopted for conducting the review process in this research. The review process was divided into six stages, namely, formulation of research questions, search strategy, study selection, quality assessment, data extraction and data synthesis. This enabled summarizing the existent literature and comprehending the research gaps. The goal is to gather empirical evidence and analyse results from existing studies to give a critically evaluated discussion on the existing trends in available

research, identify gaps in current search and provide future prospects in the area by means of answering the established research questions.

### 3.3 Findings

It was evident from pertinent literature that there has been a notable upward trend in the past decade with an increase in interest of both researchers and practitioners of using SC techniques for SA on Twitter. Since the inception of Twitter, the popularity and interest of using SA on it is discernable with the increased and focus research implementing variety of SC techniques and within diverse domains like movies, music, sports, news, health, stock exchange etc. (as shown in figure 3.1). Many reported researches were carried on the tweets fetched directly from Twitter using its API. The tweets were from a variety of domains, topics and time period (referred as topic specific or topic oriented tweets). These prominently included tweets from or about elite personalities like actors; singers; sportsperson; comedians; politicians, authors, idols; entertainers etc., news and commemoratives, health and fitness, stock market exchanges, companies like AT&T; Amazon; Apple; Google; Microsoft, consumer products like kindle, smart-phones etc., natural calamities, energy and environmental related, cyber hatred, entertainment which includes tweets about music and movies, automotive or vehicles, banking, government or public campaign or public administration, education or universities, science or technology; politics, sports, daily deals and discount, trade or commercial services or business or financial services, NGO's (charities), blogger's or journalists (freelance media professionals (FMP), taxation, terrorism etc.

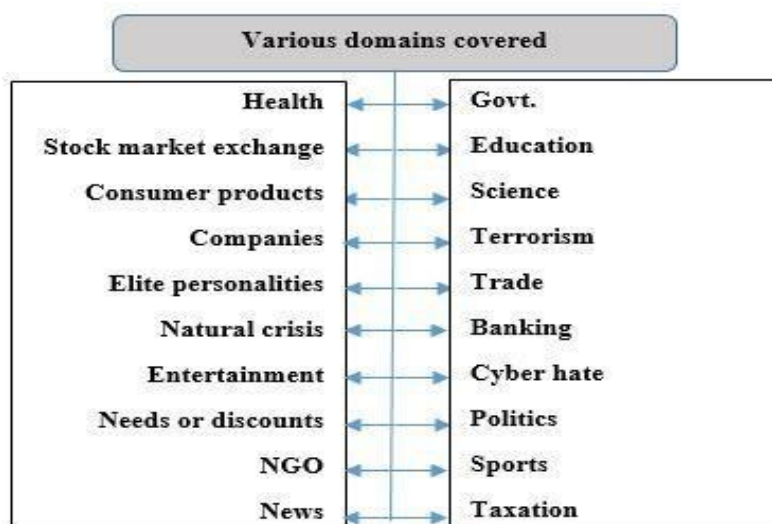
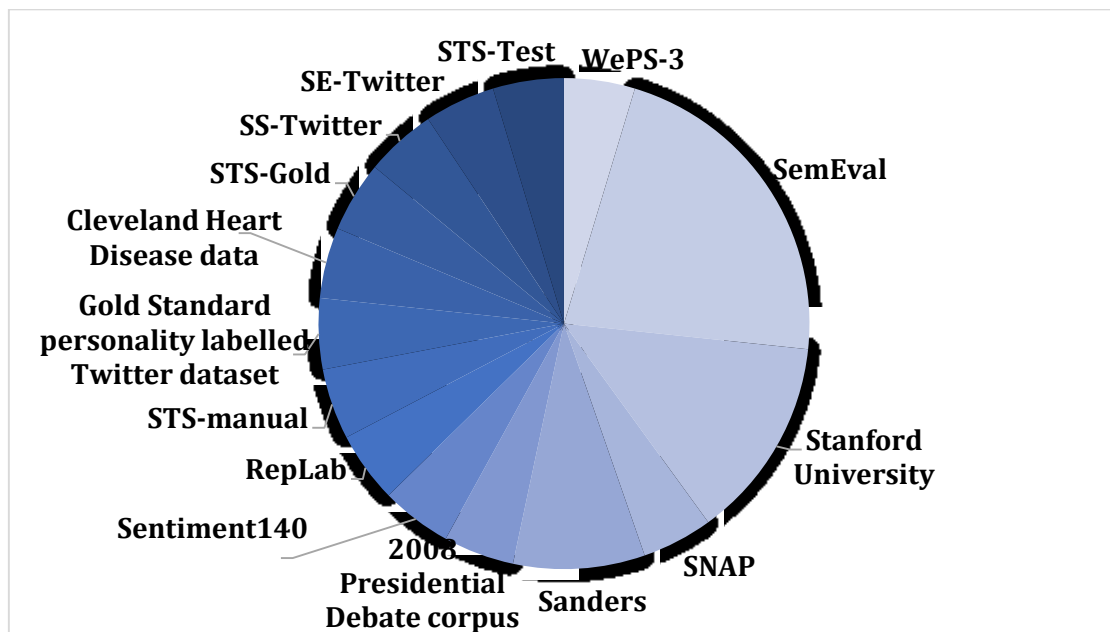


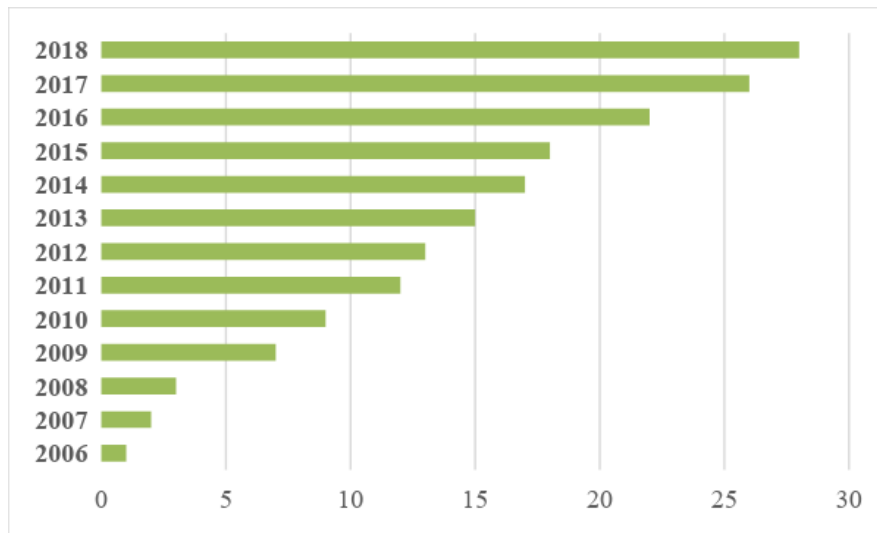
Fig. 3.1. Various domains covered for SA in past decade

Reviewing the pertinent literature it was observed that the research studies have considered a gamut of data sources including benchmark datasets as shown in figure 3.2. The probed datasets was either benchmarked or a set of random tweets collected in real-time within a selective topic/ subject/ domain. Thus, a number of datasets have been used across selected studies to conduct empirical evaluations of SC techniques for SA on Twitter. Amongst the benchmarked datasets, it was observed that SemEval datasets, especially SemEval 2007, 2013, 2014, and 2015 respectively were the most widely used ones. Next to follow was the data collected and prepared by the Stanford University and then comes the Sanders dataset. Apart from these most frequently used datasets, the other data sources that were also considerably explored are WePS-3, 2008 Presidential Debate Corpus, Sentiment140, RepLab 2012, RepLab 2013, STS-manual, Gold Standard personality labelled Twitter dataset, Cleveland Heart Disease data, STS-Gold, SS-Twitter, SE-Twitter, STS-Test.



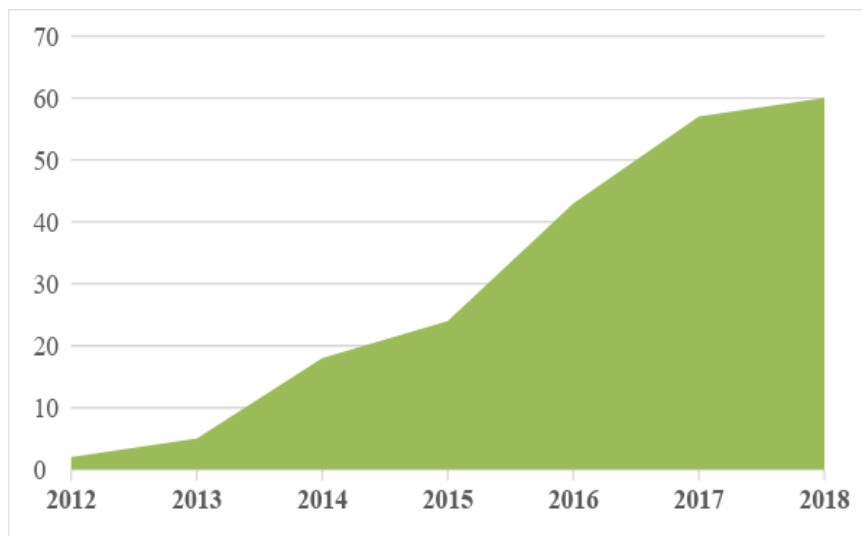
**Fig. 3.2.** Distribution of different benchmark datasets

Further with the increased trend in the usage of Twitter it was observed that random tweets on general topics from various domains have been considered for research evaluations, especially since the year 2010. The following figure 3.3 depicts the year-wise trend of published work with random tweets from various domains taken as dataset for empirical evaluation.

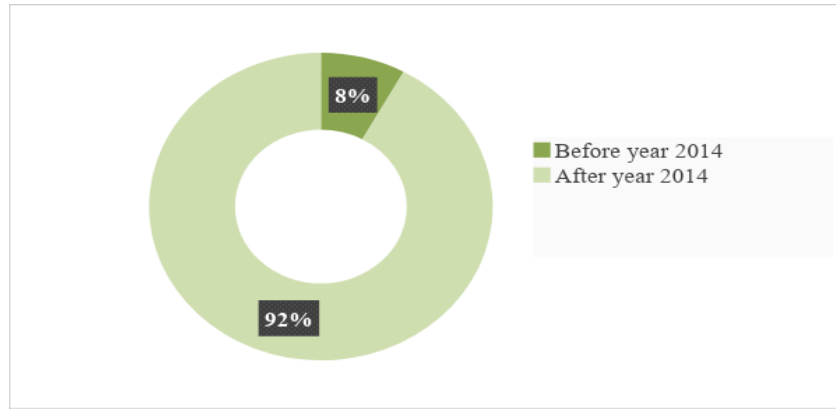


**Fig. 3.3.**Year-wise cumulative assessment of random tweets

It was also observed that the popularity of SA for Twitter using SC techniques has increased enormously after year 2014 (as shown in figure 3.4 and 3.5).



**Fig. 3.4.** Year-wise distribution of number of papers



**Fig. 3.5.** Distribution of papers after year 2014 (Number of papers/percentage of total papers)

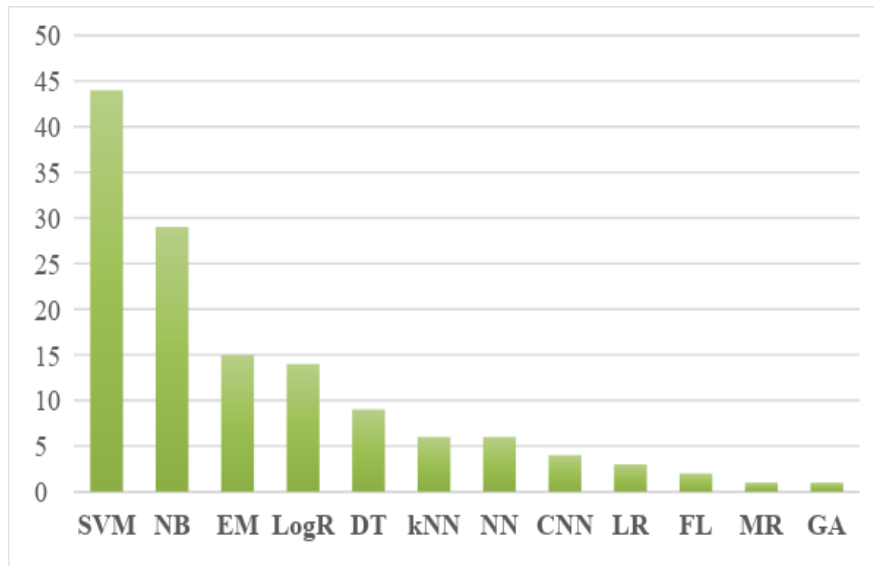
Thus, we can say that there has been a notable upward trend in the past decade, evident from studies within the domain with an increase in interest of both researchers and practitioners. Figure 3.5 is a clear indication that the popularity of SA for Twitter using SC techniques has increased enormously after year 2014, hence the future researchers and practitioners should inspect the papers being published after year 2014 so to reach to the most recent and significant research articles for the latest developments in this field. The following table 2 shows the mapping of the various SC techniques used for SA across domains in the last decade.

**Table 3.1:** Mapping of the techniques with varied domains

TECHNIQUES APPLIED	Health	Stock market exchange	ConsumerProd	Companies	Elite users	Environment/ Natural Crisis	Entertainment	Deals/discount	News	Govt.	Education	Science	Terrorism.	Trade/Finance	Banking	Cyber Hate	Politics	Sports	Taxation	NGO	FMP	Automotive
EM	✓	✓		✓	✓				✓				✓		✓	✓		✓				
DT	✓	✓		✓	✓		✓		✓		✓		✓		✓	✓	✓					✓
kNN		✓		✓	✓		✓		✓			✓					✓	✓				✓
SVM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	✓	✓	✓	✓	✓	✓
LR	✓	✓	✓	✓	✓		✓	✓	✓				✓					✓	✓			
LogR	✓			✓	✓				✓	✓	✓	✓		✓			✓	✓				
SGD			✓		✓	✓			✓													
MR		✓																				
NB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			✓	✓				✓
FL	✓		✓					✓														
NN		✓	✓	✓	✓	✓	✓	✓	✓								✓	✓				
GA	✓			✓													✓					
CNN			✓	✓	✓	✓		✓	✓													

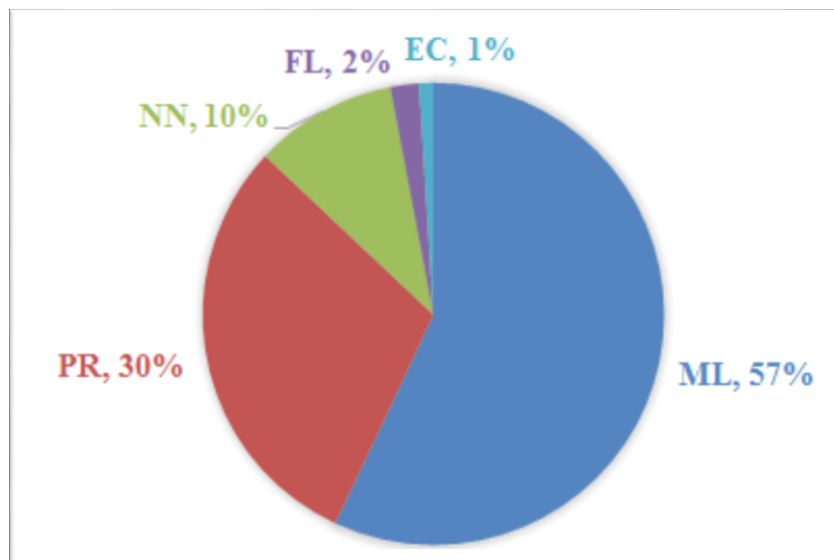


The following figure 3.6 shows the graph depicting the quantitative extent of use of various SC techniques over the past decade for SA on Twitter.



**Fig. 3.6.** Quantitative extent of use of SC techniques over past decade

SC is a broad term consisting of various techniques. Various SC techniques have been used for SA on Twitter as shown in figure 3.7 that depicts the distribution of various categories of SC techniques over the past decade.



**Fig. 3.7.** Distribution of SC techniques over the past decade (expressed in percentages)

### 3.4 Chapter Summary

This chapter discussed the convergence of Web 2.0 technologies towards sentiment analysis on social media such as Twitter using soft computing techniques. It envisages the various aspects of sentiment analysis in real life applications covering wide spectra of domains including health care, social, finance, government policies, entertainment etc. Proliferation of Web 2.0 has built strong social networks based on user's personal or professional interests. This has eventually intensified the research in the field of text mining especially sentiment analysis.

#### **Publication**

- Kumar, A. and Jaiswal, A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience*, Wiley. 2020 Jan 10:32(1):e5107. [SCIE JOURNAL, Impact factor: 1.447]. <https://doi.org/10.1002/cpe.5107>.

# Chapter 4

---

---

*Sentiment Analysis using Machine  
Learning, Swarm based Learning and  
Deep Learning*

# Chapter 4

## Sentiment Analysis using Machine Learning, Swarm based Learning and Deep Learning

Classifying sentiments in online social data is a typical natural language processing problem which takes text as input and converts these inputted texts into features that the learning algorithms can understand. SA, established as a typical text classification task [5], is defined as the computational study of people's opinions, attitudes and emotions towards an entity [6]. It offers a technology-based solution to understand people's reactions, views and opinion polarities (positive, negative or neutral) in textual content available over social media sources. Selecting features is one of the most *difficult* and imprecise part of the machine learning. The features in given data are important to the predictive models used and influence the results achieved. The quality and quantity of the features have great influence on whether the model is good or not. Though results achieved also depend on the model and the data and not just the chosen features, nevertheless choosing the right features is still very important. Better features can produce simpler and more flexible models, and they often yield better results. Given the infinite number of potential features, it's often not computationally feasible for even the most sophisticated algorithms.

The objective of this research is to process the valuable, hidden information from raw, uncertain, imprecise and high-dimensional social media data into a form more amenable to learning and maximizing predictive power using soft computing techniques. This chapter focuses on the application of baseline supervised machine learning techniques, deep learning techniques, and swarm intelligence techniques on the user-generated textual social media content for enhanced sentiment prediction.

Therefore, we seek 4.1. The methodologies and findings for the second research objective is presented in this chapter. A brief summary of above study will ends the chapter.

### 4.1 Research Objective 2

**Research objective:** To perform sentiment analysis on textual unstructured data on the Web.

## 4.2 Methodology

The methodology was primarily divided into three levels of study and implementations:

- 4.2.1 Application of Baseline Supervised Machine Learning techniques
- 4.2.2 Application of Swarm Intelligence and Machine Learning techniques
- 4.2.3 Application of Deep Learning technique

The following are the details:

### 4.2.1 Methodology 1: Application of Baseline Supervised Machine Learning Techniques

Initially, baseline supervised machine learning techniques such as naive bayesian, support vector machine, multilayer perceptron, decision tree, k-nearest neighbour, logistic regression and ensemble were analysed on social media content (real-time dataset of IMDb movie reviews, tweets and tumblogs related to four trending events (US presidential elections, Donald Trump's plans to ban Muslims from the US, Rio Olympics and release of Pokemon Go second generation) from Twitter and Tumblr, benchmark datasets named Sentiment140, SemEval 2016 & 2017 datasets) for textual sentiment prediction using the traditional, simple yet effective weighing scheme, TF-IDF.

### 4.2.2 Methodology 2: Application of Swarm Intelligence and Machine Learning techniques

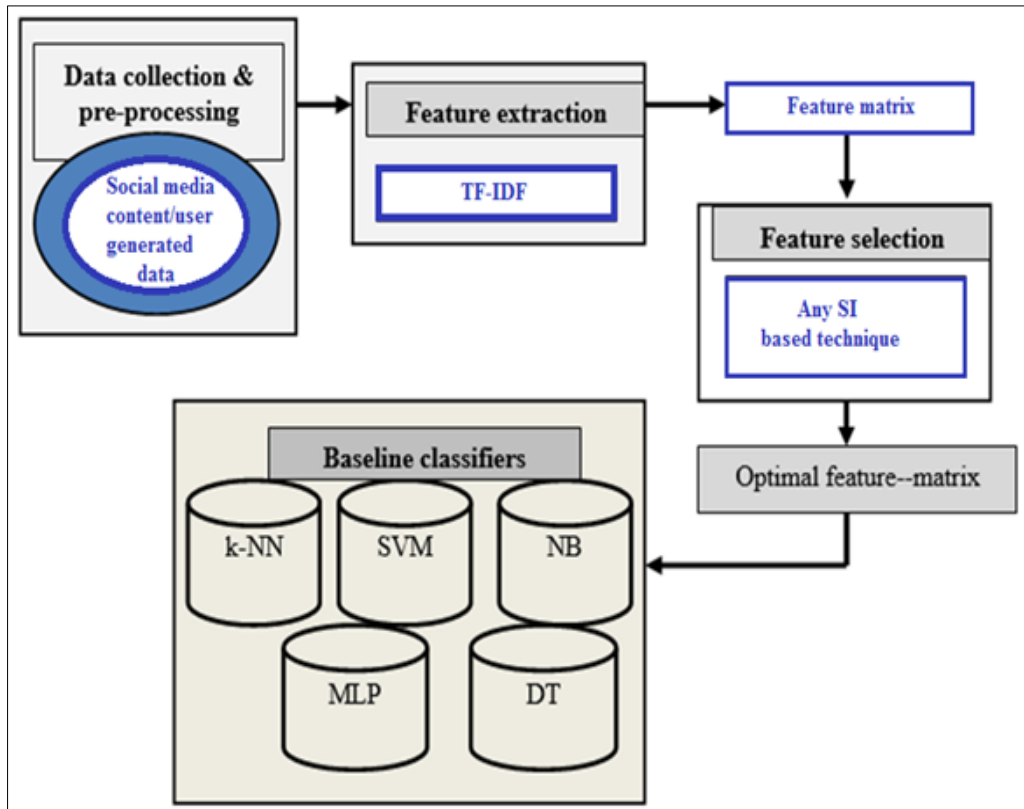
The social web data that we are discussing constantly is increasing tremendously in the recent years in form of comments, blogs, reviews and tweets etc. The nature of this data is highly un-structured and high- dimensional, making text classification a tedious task. Sentiment analysis, which is a text classification technique is applied on this data to gauge user opinion on several pertinent issues. Sentiment analysis automatically mines attitudes or views of users on specific issues. It is a multi-step process where extracting and selecting features is a vital step that controls performance of sentiment classifier. Mining and classifying incomplete and uncertain user-generated data affects the quality of sentiments derived and presents new challenges to feature selection. Moreover, social media is an informal way of communication which includes colossal usage of slangs, malformed words, short forms, colloquial expressions and mash-up words. This amplifies the ambiguity, vagueness and imprecision within the available social web content. Analysing and classifying such massive, noisy, incomplete and uncertain social web data affects the quality of sentiments derived and presents new challenges to feature selection. Past literature conforms that an optimal feature selection improves the classifier performance (in terms of speed, predictive power and simplicity of the model), reduces dimensionality,

removes noise and helps visualizing the data for model selection. Hence to increase the predictive sentiment accuracy of learned models and to improve the learning efficiency with respect to reduce storage requirements and computational cost, it is imperative to build an intelligent feature selection model. Many problems related to feature selection are NP -hard and finding an optimal subset of features is usually intractable [25]. This necessitates examining new computational methodologies for finding optimal feature set which improves the performance of the sentiment classifier in terms of predictive accuracy and result comprehensibility.

Swarm intelligence (SI) algorithms are extensively used in optimization problems. Optimization techniques could be applied to feature selection problem to produce Optimum feature set. Swarm intelligence algorithms are used in feature subset selection for reducing feature subset dimensionality and computational complexity thereby increasing the classification accuracy. One of the non-trivial sub-tasks for SA is feature selection and optimization which when implemented with swarm can lead to an enhanced sentiment prediction. Swarm intelligence algorithms are contemporary computational and behavioural metaphors for solving search and optimization problems which take collective biological patterns provided by social insects (ants, termites, bees, wasps, moths etc.) and other animal societies (fish, birds, grey wolves etc.) as stimulus to model algorithmic solutions [13].

In this research work, binary grey wolf (BGW), binary moth flame (BMF) and particle swarm optimization (PSO) were implemented with baseline machine learning techniques on the benchmark Twitter dataset (SemEval 2016, 2017) & evaluated the results using accuracy, precision and recall.

The conventional feature extraction using TF-IDF (term frequency-inverse document frequency) [87] was done on the pre-processed dataset to generate a feature matrix. Swarm-based feature optimization using the BGW and BMF (as shown in figure 4.1) is then applied independently on the generated feature matrix to acclimatize to the increased dimensionality; complexity and fuzziness in the user-generated Twitter data and eventually enhance the classifier performance. The results were empirically compared using five baseline supervised machine learning algorithms namely naïve bayesian (NB), support vector machines (SVM), multilayer perceptron (MLP), k-nearest neighbour (k-NN), and decision tree (DT). The empirical analysis demonstrates the benefit of adding swarm-based feature selection optimizer and validates that the swarm-based feature selection optimization in sentiment classification task outperforms the intrinsic TF-IDF filter based classifier.



**Fig. 4.1.** Systematic workflow for optimized sentiment classification

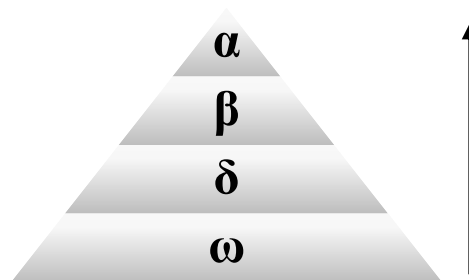
Figure 4.1 shows that the task begins by collecting data, which are tweets from benchmark datasets (SemEval 2016, 2017 datasets), annotated for polarity as positive, negative or neutral. The details about the dataset distribution has already been given in the previous sub-section. This collected data needs to be in a uniform structured format so that relevant features can be extracted. The pre-processing task includes consolidation, cleaning, transformation and reduction. A brief description about the steps involved for pre-processing has already being explained in the previous sub-section. The relevant features are then extracted using the conventional TF-IDF technique to generate the feature matrix. The optimization algorithms are then applied to this feature matrix to generate an optimal subset of features which is used to train and test the baseline ML classifiers for predicting the sentiment polarity. Results obtained are then assessed for performance accuracy.

The details about the SI algorithms applied are given as follows:

#### 4.2.2.1 Application of Binary Grey Wolf

In this sub-section, the standard grey wolf optimization (GWO) algorithm is primarily explained to enable the understanding of its binary variant. The GWO algorithm is a population-based meta-heuristic optimization algorithm which mimics the social hierarchy and the hunting behaviour of grey wolves in nature. As given by Mirjalili et al. [97], grey wolves live in a pack of 5 to 12. The pack has a hierarchical ordering of ranks; wolves of the highest rank (alpha) being revered the most.

The social hierarchy is as follows: The wolf population is divided into four categories, namely 'alpha, beta, delta and omega' as shown in figure 4.2. Alpha male or female is at the topmost position in the family pyramid and has clear dominance over other family members. This dominant wolf is the reigning leader and has authority to make leadership decisions about search spaces for hunting, resting and sleeping, and about food distribution etc. His decisions are almost obeyed by all other members in the family. Next to the alpha wolf is beta male or female wolf who is a second in command to the alpha wolf. His/her responsibility is to make sure that all the rules and regulations laid down by alpha are being followed. Beta also has the power to guide the lower level wolves in the family besides acting as a counsellor to alpha in situations when needed. Next in the hierarchy are delta wolves, which are trusted by alpha and beta. These obey and follow the decisions and orders made by alpha and beta. They majorly have the responsibility of protecting, defending and providing safety to all other wolves in the family [97]. At the lowest level of the pyramid are the omega wolves. The omega wolves always have to submit to all the other dominant wolves. They are the last wolves that are allowed to eat. They are referred to as scapegoats, and submit to all other wolves.

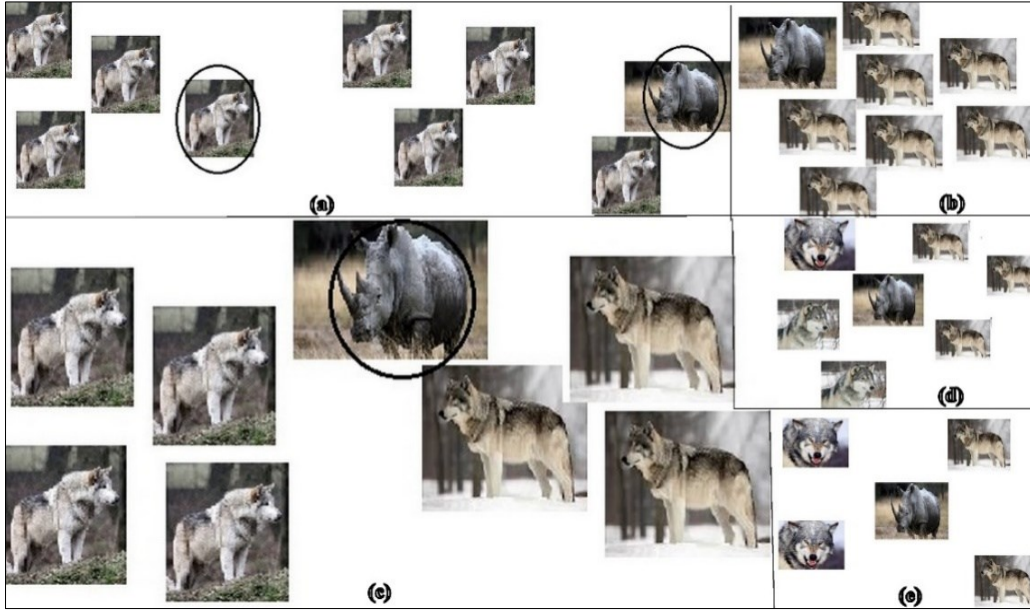


**Fig. 4.2.** BGW hierarchy (dominance increases from bottom up)

The group hunting behaviour of grey wolves includes (as shown in figure 4.3):

- Tracking, chasing, and approaching the prey
- Pursuing, encircling, and harassing the prey until it stops moving
- Attack towards the prey





**Fig. 4.3.** Hunting mechanism of Grey Wolves: (a) Tracking, chasing, and approaching the prey (b-d) Pursuing, encircling, and harassing the prey (e) attacking the prey

The optimization (hunting) model based on social hierarchy of grey wolves, the other wolves (i.e. omega) update their positions with respect to the alpha, beta and delta. That is, in the GWO algorithm, the set of possible solution sets is viewed as the wolf pack. Each solution is a wolf, and the fitness of the solution determines the position of that solution (wolf) in the hierarchy defined above. The fittest solution is regarded as the alpha ( $\alpha$ ), and the second and third regarded as beta ( $\beta$ ) and delta ( $\delta$ ), respectively. The rest of the solutions are regarded as the omegas ( $\delta$ ). The optimization problem is analogous to a hunting scenario of the wolf pack. The position of the prey is the coordinates of the most optimal solution, which is to be ultimately found. In a hunting situation, the wolves encircle the prey, and gradually close it down. The encircling behaviour is modelled using equations 4.1 to 4.4.

$$\vec{X}(t + 1) = \vec{X}_p(t) + \vec{A} \cdot \vec{D} \quad (4.1)$$

where,  $\vec{D}$  is defined below,  $t$  is the iteration number,  $\vec{A}$  and  $\vec{C}$  are the coefficient vectors,  $\vec{X}_p$  is the prey position, and  $\vec{X}$  is the

$$\vec{D} = \vec{C} \cdot \vec{X}_p(t) - \vec{X}(t) \quad (4.2)$$

The  $\vec{A}$  and  $\vec{C}$  vectors are calculated as below.

$$\vec{A} = 2a \cdot \vec{r}_1 - a \quad (4.3)$$

$$\vec{C} = 2\vec{r}_2 \quad (4.4)$$

Here,  $a$  is linearly decreased from 2 to 0 over the iterations,  $\vec{r}_1$  and  $\vec{r}_2$  are random vectors.

The hunt is guided by the alpha, and the beta and delta occasionally participate in it. The fittest solution is regarded as the alpha, the next best beta, and the third best delta. All the other solutions are called the omegas. The first three best candidate solutions guide the omegas, whose solutions are updated to get closer to the prey. The updating of the wolves positions is shown via equation 4.5.

$$\vec{X}(t + 1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \quad (4.5)$$

where,  $\vec{X}_1, \vec{X}_2, \vec{X}_3$  are defined as:

$$\vec{X}_1 = \vec{X}_\alpha - \vec{A}_1 \cdot \vec{D}_\alpha \quad (4.6)$$

$$\vec{X}_2 = \vec{X}_\beta - \vec{A}_2 \cdot \vec{D}_\beta \quad (4.7)$$

$$\vec{X}_3 = \vec{X}_\delta - \vec{A}_3 \cdot \vec{D}_\delta \quad (4.8)$$

where  $\vec{X}_\alpha, \vec{X}_\beta$  and  $\vec{X}_\delta$  are the first three best solutions in the swarm, respectively as shown via equations 6 to 8.

$\vec{D}_\alpha, \vec{D}_\beta$ , and  $\vec{D}_\delta$  are defined via equations 4.9 to 4.11.

$$\vec{D}_\alpha = \vec{C}_1 \cdot \vec{X}_\alpha - \vec{X} \quad (4.9)$$

$$\vec{D}_\beta = \vec{C}_2 \cdot \vec{X}_\beta - \vec{X} \quad (4.10)$$

$$\vec{D}_\delta = \vec{C}_3 \cdot \vec{X}_\delta - \vec{X} \quad (4.11)$$

The algorithm of GWO is given as below:

### Algorithm 1: GWO

---

#### GWO

---

**Input:**

n: Number of grey wolves in the pack

$N_{iter}$ : Number of iterations for optimization

**Output:**

$x_\alpha$ : Optimal grey wolf position

$f(x_\alpha)$ : Best fitness value

1. Initialize a population of n grey wolves' positions randomly.
  2. Find  $\alpha, \beta$  and  $\delta$  solutions based on their fitness values.
  3. While stopping criteria not met do:
    4. for each Wolf  $W_i \in$  pack do
    5. Update current wolf's position according to equation.
    6. end
-

- 
7. Update a, A, and C.
  8. Evaluate the positions of individual wolves.
  9. Update  $\alpha$ ,  $\beta$  and  $\delta$

**End**

---

The algorithm starts with initialization of n random solutions (wolves), and calculation of their fitness values. Step 2 assigns alpha, beta and delta wolves from amongst the pack, according to the fitness. Steps 3-6 mark the repeated iterations. On each iteration, for every wolf, the wolf ‘encircles’ the prey and shifts closer to it according to the equations defined above.

The GWO algorithm is a generic algorithm applicable directly on optimization (maximization) problems with continuous parameters. The feature selection problem, however, is a problem of maximizing the accuracy of classification wherein there are n parameters, corresponding to each of the n features. Each of the n parameters are discrete, assuming a value of 0 or 1, 0 implying the feature is not selected and 1 implying the feature is selected. Thus, a binary adaptation of the GWO called the binary grey wolf optimization (BGW) is used in this research. Instead of the input variables being continuous in nature, it is a binary 1xd vector, where d is the number of features out of which a subset is to be selected. The purpose of BGW optimization is to find the minimum number of features that maximize the classification performance. Thus, the fitness function is the accuracy of the classification. In this research, the findings of BGW have been observed for sentiment classification in a corpus of tweets.

In this approach the main updating equation can be formulated as shown in equation 4.12.

$$X_i^{t+1} = \text{Crossover}(x_1, x_2, x_3) \quad (4.12)$$

where  $\text{Crossover}(x, y, z)$  is suitable cross over between solutions x, y, z, and  $x_1, x_2, x_3$  are binary vectors representing the effect of wolf move towards the alpha, beta, delta grey wolves in order.  $x_1, x_2, x_3$  are calculated using equations 4.13 to 4.16 respectively.

$$x_1^d = \begin{cases} 1, & \text{if } (x_\alpha^d + bstep_\alpha^d) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4.13)$$

where  $x_\alpha^d$  is the position vector of the alpha wolf in the dimension d, and  $bstep_\alpha^d$  is a binary step in dimension d calculated as

$$bstep_\alpha^d = \begin{cases} 1, & \text{if } cstep_\alpha^d \geq rand \\ 0, & \text{otherwise} \end{cases} \quad (4.14)$$

where  $\text{rand}$  is a random number drawn from uniform distribution  $\in [0, 1]$ , and  $cstep_{\alpha}^d$  is the continuous valued step size for dimension  $d$  and can be calculated using sigmoidal function.

Similarly,

$$x_2^d = \begin{cases} 1, & \text{if } (x_{\beta}^d + bstep_{\beta}^d) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4.15)$$

and,

$$x_3^d = \begin{cases} 1, & \text{if } (x_{\delta}^d + bstep_{\delta}^d) \geq 1 \\ 0, & \text{otherwise} \end{cases} \quad (4.16)$$

The algorithm of the BGW is given as below:

---

**Algorithm 2: BGW**

---

**BGW**

---

**Input:**

$n$ : Number of grey wolves in the pack

$N_{\text{iter}}$ : Number of iterations for optimization

**Output:**

$x_{\alpha}$ : Optimal grey wolf position

$f(x_{\alpha})$ : Best fitness value

1. Initialize a population of  $n$  grey wolves at random  $\epsilon [0, 1]$
2. Find  $\alpha$ ,  $\beta$  and  $\delta$  solutions based on their fitness values.
3. While stopping criteria not met do:
4.   for each Wolf  $W_i \in$  pack do
5.     Calculate  $x_1$ ;  $x_2$ ;  $x_3$  using equations (14), (15), and (16)
6.      $x_i^{t+1} \leftarrow$  crossover among  $x_1$ ,  $x_2$  and  $x_3$
7.   end
8. Update  $a$ ,  $A$  and  $C$ .
9. Evaluate the positions of individual wolves.
10. Update  $\alpha$ ,  $\beta$  and  $\delta$

**End**

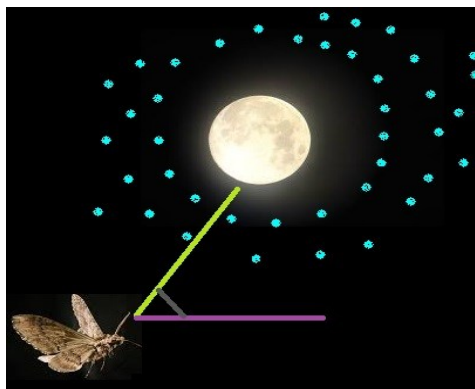
---

The algorithm starts with initialisation of  $n$  random binary solutions (wolves), and calculation of their fitness values. Step 2 assigns alpha, beta and delta wolves from amongst the pack, according to the fitness. The fitness is calculated by training the classifier according to the selected features as specified by the solution (wolf), and measuring the corresponding accuracy of the classifier. Steps 3-7 mark the repeated iterations. On each iteration, for every wolf, the wolf ‘encircles’ the prey and shifts closer

to it according to the equations defined above. The calculation of  $x_1$ ,  $x_2$ , and  $x_3$  is done as specified for the GWO.

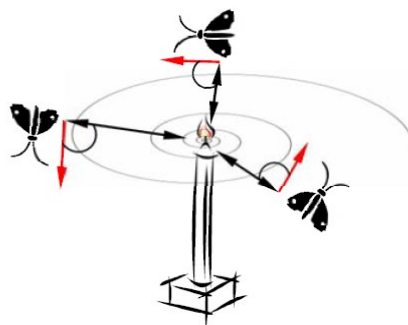
#### 4.2.2.2 Application of Binary Moth Flame

Moths are considered to have similarity with the butterfly species. They are often called as '*tiny fancy insects*'. They possess a unique navigation mechanism in which they use moon light for travelling to longer distances while remaining in a straight line. For maintaining a fixed angle with the moon light, they use '*transverse orientation*' where moths fly relative to the position of the moon [98] as shown in figure 4.4. This navigation mechanism allows the moths to move in a straight line making a fixed angle with the light of the moon.



**Fig. 4.4.** Transverse navigation mechanism of moths in moonlight

When moths encounter a man-made artificial luminous source, they attempt to maintain a constant angle with the light so as establish a straight line trajectory flight. Since this artificial source is significantly closer as compared to the moon, sustaining a constant angle to the light source results in a useless or fatal spiral fly trajectory for the moths [98] as shown via figure 4.5. It is often observed that in the case of artificial light, the moths, in due course, converge towards the luminous source.



**Fig. 4.5.** Spiral flight trajectory around close luminous sources

This behaviour of convergence of moths to such lights is being mathematically formulated to form an algorithm known as moth flame algorithm (MFO) proposed by Mirjalili [98]. It is based on the simulation of behaviour of moths and their interesting navigation techniques in the night, as explained above. It comprises of two major components i.e. *moth and flame* [98]. ‘*Moths*’ represent the actual search agents that move and travel around the search space whereas ‘*flames*’ represent the fittest or best position of the moths obtained till date. In reality, both of them are merely solutions but flames can be best understood as mid-level flags that are often dropped by these moths or agents while finding the search space. Using the transverse orientation navigation mechanism, a moth never loses its best or most optimum path (or solution) as each of the moth consistently and constantly searches around this flame or mid-level flags and consequently updates themselves to find a better solution [98]. Further, moths (Z represented in a form of matrix) can fly in any dimensional space (denoted as d), constantly changing their position vectors. The algorithm for the MFO is given below:

---

### Algorithm 3: MFO

---

#### MFO

---

**Input :** Number of dimensions, number of search agents (moths), maximum iterations

**Output :** Global best Positions of flames and moths

1. for each moth  $n_i$  ( $i = 1, \dots, m$ ) do:
  2.     for each dimension  $j$  ( $j = 1, \dots, d$ ) do:
  3.          $M(i,j) \leftarrow \text{Random}\{0,1\}$
  4.     end
  5. end
  6. for each moth  $n_i$  ( $i = 1, \dots, m$ ) do:
  7.     moth\_fitness $_i \leftarrow -\infty$
  8. end
  9. Initialize best\_flames & best\_flames\_fitness with  $M(i,j)$  and moth\_fitness respectively.
  10. While (Iteration  $\leq$  Max\_Iteration)
  11.     new\_flame\_no  $\leftarrow \text{round}(N - l*(N-1)/T)$ ;
  12.     flame\_no  $\leftarrow$  new\_flame\_no;
  13.     for each moth  $n_i$  ( $i = 1, \dots, m$ ) do:
  14.         Train classifier over training set, evaluate over testing set, store acc
  15.     end
  16.     if iteration == 1 then:
  17.          $F = \text{sort}(M)$ ;
  18.          $OF = \text{sort}(OM)$ ;
  19.     else
  20.          $F = \text{sort}(M_{t-1}, M_t)$ ;
  21.          $OF = \text{sort}(M_{t-1}, M_t)$ ;
-

---

```

22.     end
23.     r = -1 + iteration * ((-1)/Max_iteration);
24.     for each moth ni ( i = 1....n) do:
25.         for each dimension j ( j = 1....d) do:
26.             if i <= flame_no then:
27.                 D ← | Fj - Mi |
28.                 b = 1;
29.                 t = (r-1) * random() + 1;
30.                 M(i,j) = D * ebt * cos (2πt) + flame(i,j);
31.             else:
32.                 D ← | Fj - Mi |
33.                 b = 1;
34.                 t = (r-1) * random() + 1;
35.                 M(i,j) = D * ebt * cos (2πt) + flame(flame_no,j);
36.         end
37.     end
End

```

---

It starts with the initialization of the position of the moths in a matrix, using the function I described in equation 4.17, as shown in Step 1-5.

$$MFO = (I,P,T) \quad (4.17)$$

This equation signifies that the MFO Algorithm is a tuple of three values that approaches the global optimal of the optimization problems.

$$I : \Phi \rightarrow \{M, OM\} \quad (4.18)$$

I is a function that institutes a random population of moths and their corresponding fitness values. The above equation 4.18 depicts the methodical model of this function. Steps 9-35 mark the actual implementation of the said algorithm.

A loop from steps 10-36 is executed, which in each iteration, updates the number of flames according to equation 4.24. Thereafter the matrix containing fitness values of moths is computed using fitness function given by equation 4.25.

$$P : M \rightarrow M \quad (4.19)$$

The P function is the principal function that is responsible for the movement of the moths around the search space. It receives the matrix M, and returns its updated version, as shown via equation 4.19.

Steps 16-19 are executed for the 1<sup>st</sup> iteration, when we don't have any past populations of moths. Here, the moths are sorted according to their fitness values from best to worst. The flame matrix is thereby updated with these values. Steps 17-20 depict the case when a past population of moths exist, and can be used in updating of flame matrix. The previous and current populations are merged, and then sorted according to fitness values. Subsequently, updating are made on the flame matrix. Steps 21 to 35 depict the updating of moth positions using equation 4.21 and 4.22. The convergence constant  $r$ , and parameter  $t$ , is accordingly calculated as shown via equation 4.20.

$$T : M \rightarrow \{\text{true, false}\} \quad (4.20)$$

The  $T$  function returns true if the termination norm is satisfied and false if termination norm isn't satisfied.

$$M_i = S(M_i, F_j) \quad (4.21)$$

$M_i$  indicates the  $i$ th moth,  $F_j$  indicates the  $j$ th flame, and  $S$  is the spiral method, as shown via equation 4.21 and 4.22.

Any types of spiral can be utilized here subject to the following conditions:

- Spiral's initial point should start from the moth.
- Spiral's final point should be the position of the flame.
- Fluctuation of the range of spiral should not exceed the search space.

$$S(M_i, F_j) = D_i \cdot e^{bt} \cdot \cos(2\pi t) + F_j \quad (4.22)$$

This is the definition of the logarithmic spiral is given via equation 4.23.

$$D_i = |F_j - M_i| \quad (4.23)$$

Here  $D_i$  defines the distance of the  $i$ th moth for the  $j$ th flame,  $b$  is a constant that explains the shape of the logarithmic spiral, and  $t$  is a rand no. in  $[-1,1]$ , as shown via equation 4.24.

$$N_{\text{flames}} = \text{round} ( N - l * (N-1) / T ) \quad (4.24)$$

Here,  $l$  is the current number of iteration,  $N$  is the maximum no. of flames, and  $T$  is maximum no. of iterations, as shown via equation 4.25

$$f_{\theta} = \alpha \cdot E + (1-\alpha) * (\sum \theta_i) / N \quad (4.25)$$



Here,  $f_{\theta}$  is fitness function given a vector  $\theta$  sized  $N$  with 0/1 constituents depicting unselected/selected features,  $N$  is total no, of features in data set,  $E$  is classifier error rate and  $\alpha$  is constant manipulating the significance of classification performance to number of features selected.

To facilitate the task of feature selection, a binary version of moth flame optimization (MBF) algorithm has been designed. Standard MFO provides continuous valued positions of solutions, which are, flames and moths. Using a sigmoid function, the continuous valued method is transformed into discrete binary values. A solution binary vector is used, where 1 represents feature selected and 0 represents otherwise. A threshold is applied to convert the value output by sigmoid function into binary values, in other words, discretize it.

The algorithm of BMF is shown below:

#### Algorithm 4: BMF

---

##### BMF

---

**Input:** Number of dimensions, number of search agents (moths), maximum iterations

**Output :** Global best positions of flames and moths

1. for each moth  $n_i$  ( $i = 1, \dots, m$ ) do:

2.

.

.

.

36.  $\text{sigma} \leftarrow \text{random}(0,1)$

37. if  $\text{sigma} < S(M(i,j))$  then:

38.  $M(i,j) \leftarrow 1$

39. else:

40.  $M(i,j) \leftarrow 0$

41. end

42. end

Steps 1-35 are same as given in the MFO

**End**

---

The steps 1-35 of the pseudo-code are the same as that of the generic (continuous) MFO algorithm. Steps 36-40 depict the binary version of MFO introducing an auxiliary function which compares the values obtained from the generic algorithm to a certain threshold, and thereby assigns binary values. A conditional operator is used to branch 2 different paths for the binary values.

The function  $S$  is called sigmoid function, which transforms the continuous valued positions provided by standard MFO Algorithm into discrete binary values, as shown via equations 4.26 and 4.27.

$$S(xi(t)) = \frac{1}{1 + e^{-xi(t)}} \quad (4.26)$$

$$M(i,j) = \begin{cases} 1, & S(M(i,j)) > \sigma \\ 0, & \text{otherwise} \end{cases} \quad (4.27)$$

$\sigma \sim U(0,1)$ .

The constants and parameters used are as follows:

- b: constant defining the shape of the logarithmic spiral
- r : convergence constant, decreases linearly from -1 to -2 over the course of iteration
- t : parameter defining how close should the next position of moth be to flame
- $\sigma$ : sigma, random number generated from uniform distribution between 0 and 1.

### 4.2.2.3 Application of Particle Swarm Optimization

In this research, we exhibited the use of PSO for enhanced SA of tweets. It was given by Kennedy & Eberhart [99-100] emulating the behavioral aspects of flock of birds etc. The initial PSO version holds these initial imitations. Post this, concept of ‘inertia weight’ was given by Shi et al. [101] for demonstrating the basic PSO.

This algorithm consists of random set of possible solutions, commonly known as ‘particles’, that is often represented as a point in “S-dimensional search space”. The concept of a particle is also displayed via figure 4.6 where the i-th particle is denoted by  $X_i = (X_{i1}, X_{i2}, X_{i3}, \dots, X_{iS})$ .

Fitness values [101] are being associated with these particles that are assessed using a fitness function that needs to be further optimized for yielding more efficient results. With each iteration, a particle always updates its “pbest and gbest” values. Former represents the best fitness value achieved by the particle i.e.  $P_i(t)$ , represented as  $P_i = (P_{i1}, P_{i2}, P_{i3}, \dots, P_{iS})$ , whereas the latter denotes the best value attained by any particle in the population denoted by g(t).

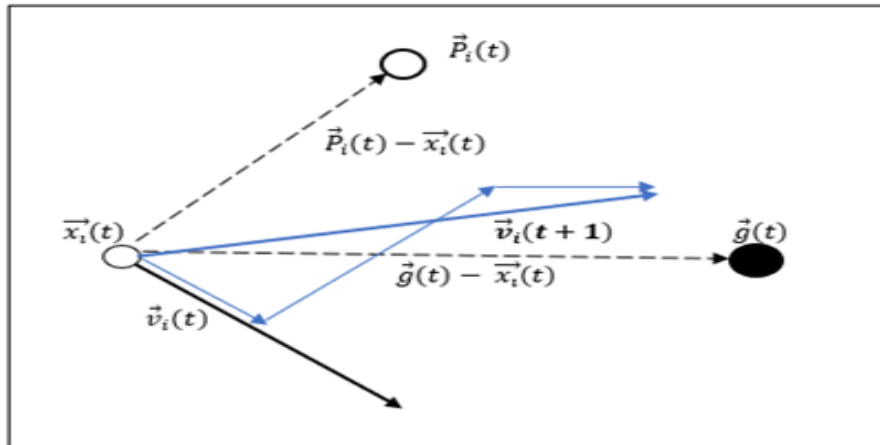


Fig. 4.6. The concept of a flying particle.

Algorithm of PSO is shown below:

**Algorithm 5: PSO**

**PSO**

Input:  $r_1$  and  $r_2$  are uniformly distributed random functions

$r_1$  and  $r_2 \in [0,1]$

$d = 1, 2, 3, \dots, S$  is the dimension

$w$  is inertia weight

$c_1$  and  $c_2$  are the acceleration coefficients

FOR each particle  $i$

FOR each dimension  $d$

Initialize position  $X_{id}$  randomly within range

Initialize velocity  $V_{id}$  randomly within range

End FOR

END FOR

Iteration  $t = 1$

DO

FOR each particle  $i$

Calculate Fitness Value

IF Fitness ( $X_{id}(t)$ ) > Fitness ( $P_{id}(t)$ )

$P_{id} = X_{id}$

End IF

End FOR

Chose the particle having the best Fitness value as the  $g_d(t)$

FOR each particle  $i$

FOR each dimension  $d$

Calculate velocity according to the equation

$$V_{id}(t+1) = w \cdot V_{id}(t) + r_1 \cdot c_1 \cdot (P_{id}(t) - X_{id}(t)) + r_2 \cdot c_2 \cdot (g_d(t) - X_{id}(t))$$

---

Update particle position according to the equation

$$X_{id}(t + 1) = X_{id}(t) + V_{id}(t + 1)$$

END FOR

END FOR

t = t + 1

WHILE maximum iterations or minimum error criteria are not attained

---

### 4.2.3 Methodology 3: Application of Deep Learning Technique

The key challenges pertaining to a sentiment classification task are related to the characteristics of real-time datasets which are inherently ‘imbalanced’, ‘heterogeneous’, ‘multimodal’ and ‘cross-lingual’. Deep learning (DL) techniques have achieved state-of-the-art results owing to their power of self-tuning, the competence, and learning skills by generalizing from the training data, hierarchical learning capabilities and generalization [83]. Deep learning methods aim at learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower level features. Automatically learning features at multiple levels of abstraction allow a system to learn complex functions mapping the input to the output directly from data, without depending completely on human-crafted features. Hence, deep learning automatically finds out the features which are important for classification. Deep learning algorithms try to learn high-level features from data. This is a very distinctive part of Deep Learning and a major step ahead of traditional Machine Learning. Therefore, deep learning reduces the task of developing new feature extractor for every problem. A deep learning algorithm takes a long time to train. This is because there are so many parameters in a deep learning algorithm that training them takes longer than usual. In this research, we implemented convolution neural network (CNN) using GloVe word embedding for sentiment analysis on Twitter data.

## 4.3 Findings

### 4.3.1 Findings of Methodology 1

Above specified datasets were analysed for textual sentiment analysis using baseline ML techniques. The findings were as follows:

#### 4.3.1.1 Findings for Twitter and Tumblr Datasets

This sub-section discusses the results obtained by application of baseline ML techniques on Twitter-Tumblr dataset. Desired data about the four top-most trending events from Twitter and Tumblr using their respective APIs was collected. We considered the four most trending topics (over the same duration, to maintain uniformity) in years 2016-2017, that is, the US presidential elections (2016), Donald Trump’s plans to ban muslims from the US (2017), rio olympics (2016) and release of pokemon Go second generation (2017) and extract nearly 3000 tweets and 3000 tumblogs. The next step was to pre-

process the dataset by removing any noise, incompleteness, inconsistency within it. The data was firstly cleaned by removing #tags, @symbols, URLs, email ids, punctuations, symbols, numbers, digits, alphanumeric, non-English posts and then transformed removing stop-words, expanding short-forms & replacing emoticons, tokenization and stemming. After pre-processing, 2,272 relevant tweets and 1,983 relevant tumblogs were obtained (removing redundant, junk tweets & tumblogs). The pre-processed tweets & tumblogs were manually labelled (annotated) to accomplish coarse grain sentiment analysis (positive, negative or neutral). It was then assessed using five supervised soft computing techniques (as explained in table 4.1) namely, naive bayesian (NB), support vector machine (SVM), multilayer perceptron (MLP), decision tree (DT) and k-nearest neighbour (kNN) in Weka.

**Table 4.1:** Details about SC techniques used

Techniques	Description
<b>Naive Bayesian (NB)</b>	These methods are based on applying Bayes theorem with the “naive” assumption of independence between every pair of features [27, 30-32]. They require a small amount of training data to estimate the necessary parameters.
<b>Support Vector Machine (SVM)</b>	Formally, a SVM constructs a hyper-plane or set of hyper-planes in infinite-dimensional space, which can be used for classification, etc. Intuitively, a good separation is achieved by the hyper-plane that has the largest distance to the nearest training-data point of any class. In general, the larger the margin the lower the generalization error of the classifier [27-36].
<b>Multilayer Perceptron (MLP)</b>	They belong to the class of ‘Feed Forward Artificial Neural Network’ having input layer, one or more hidden layers and an output layer [84]. The leftmost layer (input) consists of a set of neurons representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation followed by a non-linear activation function, for example like the hyperbolic tan function etc. The output layer receives the values from the last hidden layer and transforms them into output values.
<b>Decision Tree (DT)</b>	This algorithm breaks down a dataset into smaller and smaller subsets and simultaneously develops an associated DT, having decision nodes and leaf nodes. DT using J48 is the implementation of algorithm ID3 (Iterative Dichotomiser3) developed by the Weka project team which is a top-down, greedy search through the space of possible branches with no backtracking [39, 44].
<b>k-Nearest Neighbours (kNN)</b>	It is non-parametric method used for classification where input consists of the k- closest training examples in the feature space. In case

	of classification, the output is a class membership [34]. An object is classified by majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small [34, 49]).
--	---

The training and the testing dataset selection procedure has been done using 5-fold cross validation method. The results were evaluated based on efficacy measures like precision, recall, accuracy for probing the capabilities and scope of sentiment analysis within the two micro-blogs. Accuracy [29] is defined as proximity of a measurement to its true value. It is calculated as a proportion of true positives and true negatives among total inspected cases. Precision [28-29] defines the exactness of any classifier. A higher precision value indicates fewer ‘false positives’ and vice versa. It is given as the ratio of true positives to all the predicted positives. Recall [29] defines the sensitivity or the completeness of any classifier. A higher recall value indicates less ‘false positives’ and vice versa. Recall and precision are bounded by inverse relation with each other.

The results of applying the aforesaid techniques on the data obtained from both Twitter and Tumblr on the four trending topics are presented in the tables 4.2 to 4.5 using KPIs (expressed in percentages).

**Table 4.2:** Summary of results for ‘RIO OLYMPICS’

Techniques	Twitter			Tumblr		
	P	R	A	P	R	A
<b>NB</b>	74.5	73.5	73.9	60.0	59.3	59.3
<b>SVM</b>	70.0	70.5	70.5	71.7	65.8	65.8
<b>MLP</b>	67.4	68.2	68.2	69.9	64.9	64.9
<b>DT</b>	60.7	61.4	61.4	72.3	58.5	58.45
<b>kNN</b>	62.8	62.9	62.9	67.4	61.2	61.24

**Table 4.3:** Summary of results for ‘Release of Pokemon Go Gen’

Techniques	Twitter			Tumblr		
	P	R	A	P	R	A
<b>NB</b>	59.5	58.6	58.6	68.2	70.2	70.17
<b>SVM</b>	62.0	60.4	60.4	76.6	72.5	72.5
<b>MLP</b>	59.1	57.2	57.2	74.2	72.5	72.47
<b>DT</b>	52.6	53	52.9	44.7	66.9	66.9
<b>kNN</b>	54.1	54.5	54.5	71.4	71.3	71.35

**Table 4.4:** Summary of results for ‘US presidential elections’

Techniques	Twitter			Tumblr		
	P	R	A	P	R	A
<b>NB</b>	59.2	60.2	60.3	62.6	64.8	64.75
<b>SVM</b>	67.1	64.7	64.7	73.8	70.8	70.75
<b>MLP</b>	59.1	58.4	58.4	69.0	68.9	68.9
<b>DT</b>	57.4	58.0	57.9	49.4	58.0	57.9
<b>kNN</b>	57.56	57.2	57.2	64.9	64.5	64.49

**Table 4.5:** Summary of results for ‘Donald Trump’s claims of Muslim Ban’

Techniques	Twitter			Tumblr		
	P	R	A	P	R	A
<b>NB</b>	72.9	72.5	72.5	68.8	69.5	69.53
<b>SVM</b>	76.2	74.5	74.5	78.2	69.5	69.5
<b>MLP</b>	65.7	67.8	67.8	68.6	68.8	68.75
<b>DT</b>	66.1	65.8	65.8	67.8	62.5	62.5
<b>kNN</b>	60.8	61.1	61.1	62.9	64.8	64.8

Figure 4.7 depicts the results obtained by application of baseline ML techniques on Twitter-Tumblr dataset.

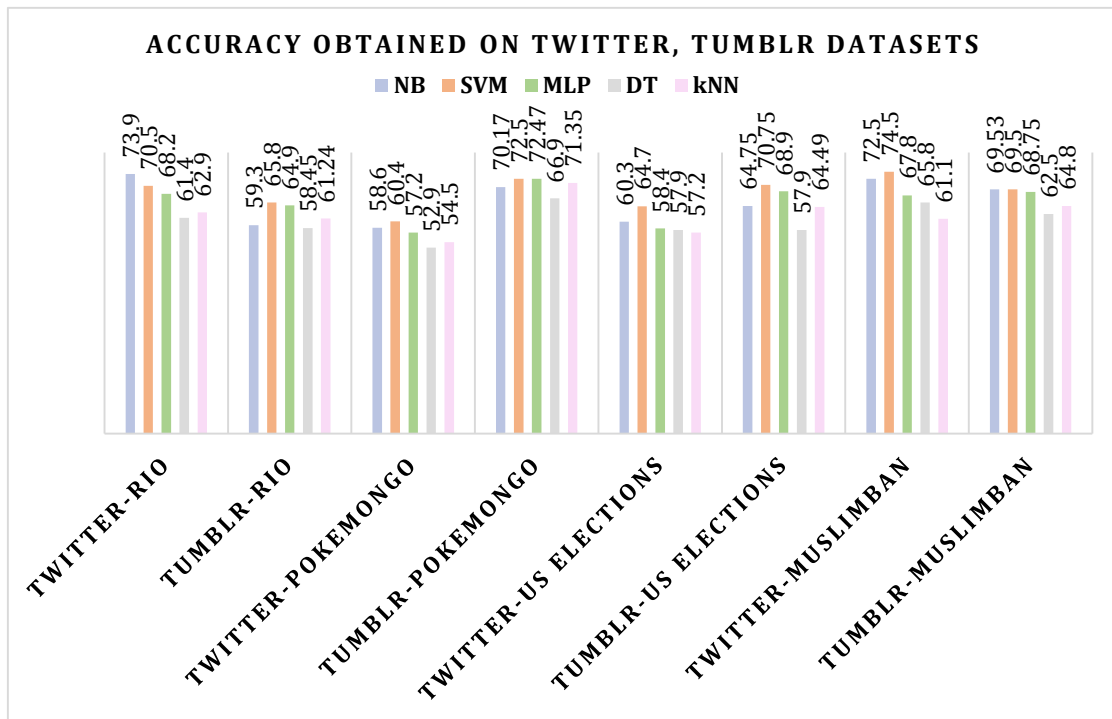


Fig. 4.7. Accuracy results on using ML techniques on Twitter, Tumblr

The results of our study suggests that the best sentiment accuracy and precision is achieved using support vector machine (SVM) for both Twitter and Tumblr. SVM outperformed all other supervised classification algorithms in terms of P and A, followed by naïve bayesian (NB) and k-nearest neighbour (k-NN) techniques, with the highest P for SVM, (approximately 76% for Twitter and approximately 78% for Tumblr). Multi-layer perceptron neural network (NN) also showed encouraging results, quite akin to NB for both the micro-blogs. Next to NN, decision trees (DT) had comparable accuracy. The observed variation in the results is merely due to the fact that large chunks of tweets and tumblogs were skewed towards negative sentiments, predominantly for the topics like rio olympics, US presidential elections and Donald Trump’s claim for muslim ban. From the results it is deduced that improved and optimized results were observed for Tumblr in contrast to Twitter.



#### 4.3.1.2 Findings for SemEval 2016, 2017 Datasets

Two benchmark Twitter datasets were used. These are: SemEval 2016 (Task 4, subtask-A) and SemEval 2017 (Task 4, subtask-A). The selected datasets consists of ‘tweet id’s’ which were annotated as positive, negative and neutral. The SemEval 2017 dataset comprises of around 5742 neutral, 2352 positive and 3811 negative tweets, whereas the SemEval 2016 dataset consists of 10341 neutral, 7059 positive and 3231 negative tweets. Few of the tweets were not available for download, some were removed and few had an altered privacy status that eventually left us with approximately 5000 tweets each for the datasets. These were then pre-processed. Pre-processing was done to clean and transform the data for relevant feature extraction [25]. It involves cleaning the dataset from noise. Noise here basically connotes the language irregularities often present in any micro-blog text. As the noisy and unstructured data affects the quality of the sentiment classification task, the data is converted to structured input format [4, 23].

The pre-processing procedure employed in this study involves the following tasks:

- Removal of all the URL links, non-ASCII characters, non-English characters, numbers, stop words like ‘the, at, as, of, or, to’ etc.
- Transforming any negative mentions to their original meaningful words by using Internet Slang Dict<sup>1</sup>. For example replacement of “couldn’t to could not” etc., expansion of acronyms and slangs etc.
- Replacing the emoticons to their original textual forms by using emoticon dictionary.<sup>2</sup>
- Stemming and tokenizing. Stemming is done on text in order to preserve the root of the word, for example it reduces helping to its root word i.e. help. Tokenization is defined as process of chopping or splitting any text into “individual words or sequences of words (n-grams)”. Tokenization was done using the Tweet-NLP [86].

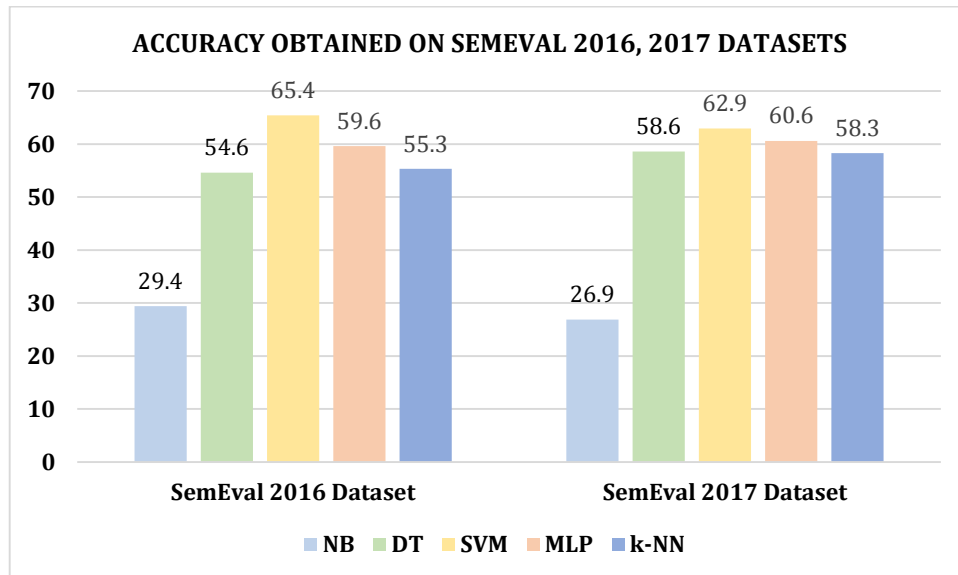
Then, feature extraction and selection was done using term-frequency-inverse document frequency (TF-IDF) [87] which was used to generate the feature matrix. This numeric statistic depicts the relevance of a word in the selected corpus. Thereafter, five baseline classifiers namely, NB, SVM, DT, k-NN and MLP have been modelled. A brief description of these techniques was already given in above sub-section. Python was used for implementation of all the selected algorithms. Python was chosen because of its suitability to intense mathematical programming, huge community support and numerous open source packages available. Various python libraries used were scikit-learn (it has been used to implement various supervised classification algorithms on which the accuracy is determined after feature selection), numpy (this library has mathematical functions that has support to work with large arrays and matrices. It is optimized to reduce the computation time required to work with large matrices. Its implementation involved large matrices for extracting features with and without TF-IDF), scipy (it is the library which is used for scientific computing, optimization etc.) and

---

<sup>1</sup> <https://www.noslang.com/dictionary>

<sup>2</sup> [http://en.wikipedia.org/wiki/List\\_of\\_emoticons](http://en.wikipedia.org/wiki/List_of_emoticons)

NLTK3 (it is a very comprehensive collection of programs and libraries aimed at statistical and symbolic natural language processing for English language. It has lot of inbuilt functions like Stemmers, Lemantiser, and Tokeniser etc.). Ten-fold cross validation method was used. Figure 4.8 depicts the results obtained by application of baseline ML techniques on SemEval 2016 & 2017 datasets.



**Fig. 4.8.** Accuracy results on using ML techniques on SemEval 2016, 2017

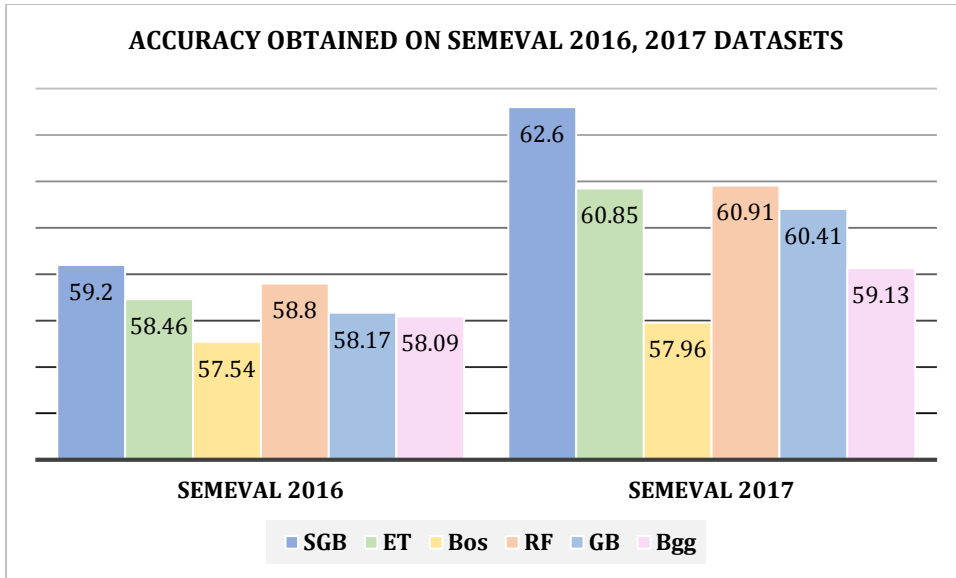
From above figure, we observe that support vector machines (SVM) had attained the highest accuracy in the range of 63% to 65% whereas naïve bayesian (NB) reported the lowest accuracy of around 50% to 55% when applied to SemEval 2016, 2017 [88-89] datasets as depicted in figure 15. We could also deduce that the multilayer perceptron (MLP) shows encouraging results by producing accuracy of around 60% to 61% for SemEval 2016, 2017 datasets respectively. Next to MLP, comes decision tree (DT) that shows accuracy within the range of 55% to 59% for SemEval 2016, 2017 datasets. After this, k-nearest neighbor (kNN) follows which yielded accuracy of approximate 55% and 58% for SemEval 2016, 2017 datasets.

Figure 4.9 depicts the results obtained by application of ensemble techniques, namely, random forests (RF), bagging (bagg), boosting (Bos), gradient boosting (GB), stochastic gradient boosting (SGB) and extra trees (ET) on SemEval 2016 & 2017 datasets. A brief description about these ensemble techniques is provided in table 4.6.

<sup>3</sup> Natural Language Toolkit: <https://www.nltk.org/>

**Table 4.6:** Description of ensemble techniques

<b>Technique</b>	<b>Description</b>
<b>Bgg</b>	Also called as Bootstrap Aggregating. It is a combination of bootstrapping and aggregation. It explicates the formation of decision trees for each of the boot strapped sub-samples [90-92]. Post this; most efficient predictor is formed using an algorithm that aggregates these decision trees.
<b>RF</b>	This technique introduces randomization into the tree building process by randomly selecting a subset of features instead of selecting all the features in order to produce better and stable results [93]. It is also called as extended bagged decision trees.
<b>Bos</b>	For Bos, Adaboost algorithm commonly known as adaptive boosting was employed. Here, weighted majority vote principle is used for classification [90]. It envisages the training of the base learners on the weighted data in sequence.
<b>GB</b>	Also called as gradient tree boosting. It includes different loss functions that need to be minimized at each stage in a sequential way [94].
<b>SGB</b>	Also called as gradient boosting machines. It is a kind of modification of boosting procedure which involves the random extraction of the sub-sample of training corpus for every iteration without any replacement from the main corpus. This random selection procedure is then used for fitting the base learner [95].
<b>ET</b>	It is another variation of bagging that involves building of random trees from training corpus samples [96]. Instead of using the bootstrap samples, it randomly selects features at each node.



**Fig. 4.9.** Accuracy results on using Ensemble techniques on SemEval 2016, 2017

Tables 4.7 shows the summarized results obtained by application of ensemble techniques on SemEval 2016, 2017 datasets when evaluated using A, P and R.

**Table 4.7:** Summary of results for application of ensemble techniques on SemEval 2016, 2017

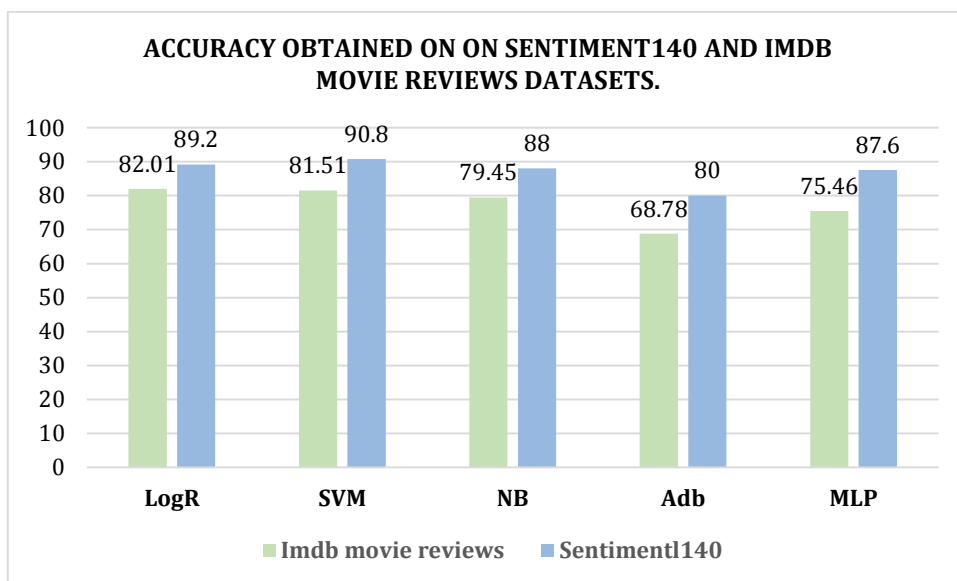
Techniques	SemEval 2016			SemEval 2017		
	P	R	A	P	R	A
<b>SGB</b>	58.26	59.20	59.20	62.31	62.60	62.60
<b>ET</b>	57.14	58.46	58.46	61.71	60.85	60.85
<b>Bos</b>	56.41	57.54	57.54	57.63	57.96	57.96
<b>RF</b>	58.07	58.80	58.80	60.32	60.91	60.91
<b>GB</b>	56.06	58.17	58.17	59.99	60.41	60.41
<b>Bgg</b>	56.06	58.09	58.09	58.70	59.13	59.13

The results depicts that the maximum A is attained by SGB on both the datasets, i.e. around 59% and 63% respectively. Amongst all, Bos shows the minimum accuracy of around 57% for both the datasets. After SGB, RF shows an appreciable classification accuracy of around 58% for SemEval 2016 and 61% for SemEval 2017, followed by ET, GB and Bgg. It was observed that the results of ET and GB are comparable for both the datasets.

### 4.3.1.3 Findings for IMDb movie review & Sentiment140 Datasets

Sentiment analysis is quite imperative and serve as info-foundation for smart cities as they have the ability to harness the opinions or sentiments accurately based on the computation technology applied. In this research, we have also, demonstrated the real-life application of sentiment analysis on online movie reviews and how it could be utilized for potential decision-making. The evaluation is done on two different datasets of Twitter. One of the dataset is a benchmark corpus named 'sentiment140' and the other is real-time dataset of IMDb movie reviews. Former dataset consisted of 1600000 tweets and the latter dataset consisted of 25000 tweets. These tweets were then subjected to the process of data cleaning, also called as pre-processing (as discussed in previous sub-section). TF-IDF vectorization [87] was used for feature selection. Post this, supervised learning techniques namely logistic regression (LogR), naïve bayesian (NB), adaboost (Adb), multi-layer perceptron (MLP) and support vector machines (SVM) were applied using python libraries (a brief description about these techniques has already been given in previous sub-section).

Below graph in figure 4.10 depict the results obtained by application of supervised techniques on Sentiment140 and IMDb movie reviews datasets.



**Fig. 4.10.** Accuracy results on Sentiment140 and IMDb movie reviews.

Above results show that the final output of the processes yields purposeful information for these datasets. It also depicts that amongst all movie reviews, majority of the reviews were towards higher range of viewer satisfaction i.e. within the range of approximately 68% to 82%. This higher level shows these reviews were correct and absolute that could be utilized by the viewers for judicious decision making, whereas there were around 32% to 18% that happened to be quite inaccurate in sentiment prediction. It all depends on the technique employed. Different techniques yield different

results based on their computational abilities. Amongst all, LogR predicted sentiments most aptly, i.e. it was able to accurately predict 82% of the sentiments as positive, negative or neutral and for rest of the cases, it failed to predict absolutely. Likewise, for the other dataset, different techniques gave different results for sentiment prediction about the opinions for any topics, brand etc. Highest accuracy of around 91% was obtained using support vector machines that mean that for 91% of the times, it was able to correctly classify the sentiment as either positive or negative about any brand or topic etc., and for rest of the cases, it went out inaccurate. Consequently, these techniques showed different sentiment prediction rate owing to their respective competence and ability. The techniques that are showing higher accuracy percentages offer better credibility to the users.

### 4.3.2 Findings of Methodology 2:

Above specified datasets were analysed for textual sentiment analysis using above specified SI+ ML techniques. The findings are as follows:

#### 4.3.2.1 Findings of BGW and BMF for SemEval 2017 Dataset

The following table 4.8 depicts the accuracy achieved by baseline classification techniques without and with optimization for the SemEval 2017 dataset.

**Table 4.8:** Accuracy gain using BGW and BMF in SemEval 2017

<b>Technique</b>	<b>Non-optimized approach (TF-IDF) Accuracy (%)</b>	<b>Optimized approach (TF-IDF + BGW) Accuracy (%)</b>	<b>Increase in accuracy (%)</b>	<b>Optimized approach (TF-IDF + BMF) Accuracy (%)</b>	<b>Increase in accuracy (%)</b>
<b>NB</b>	26.9	36	9.1	39.9	13
<b>DT</b>	58.6	69.3	10.7	67.7	9.1
<b>SVM</b>	62.9	74.5	11.6	73.8	10.9
<b>MLP</b>	60.6	69.6	9	70.9	10.3
<b>k-NN</b>	58.3	67	8.7	69.2	10.9

Here, best accuracy with BGW optimization is achieved by SVM, i.e. 74.5% and NB shows the lowest accuracy of around 36%. The maximum accuracy gain was obtained by SVM (11.6%) while DT showed an appreciable 10.7% gain in accuracy followed by approximately 9% gain for NB and MLP and k-NN. The average accuracy gain using BGW on SemEval 2017 dataset was of 9.82%.

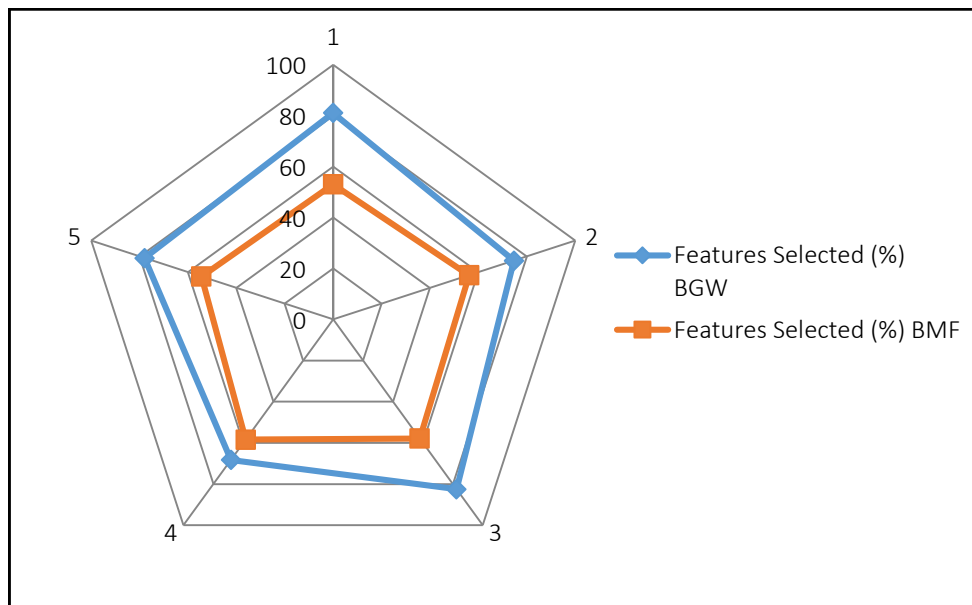
The results also indicate that the best accuracy with BMF optimisation is again achieved by SVM on SemEval 2017 dataset, i.e. 73.8%. NB showed the lowest accuracy of around 39.9% using BMF optimizer.

Again, the maximum accuracy gain was observed for NB (13%) while SVM and k-NN showed an equally appreciable 10.9% gain in accuracy followed by approximately 10% gain for DT and MLP. The average accuracy gain using BMF on SemEval 2017 dataset was of 10.84%.

The following table 4.9 displays the number of features selected in SemEval 2017 dataset using the two optimizers. In non-optimised approach all the classification algorithms used the same number of features, which is 2658. After applying BGW for feature selection the minimum number of features selected were 1814 (MLP) which is 68.24 % selection and maximum were 2192 (SVM) which is 82.46% selection. The analysis shows that on an average 76.876% features were selected. Using BMF the minimum number of features selected were 1411 (NB) that is, 59.91% features selected and the maximum were 1552 (MLP) which is 58.38% features selected. The results show that on an average 55.99% features were selected, as depicted via figure 4.11.

**Table 4.9:** Percentage of features selected using BGW and BMF in SemEval 2017

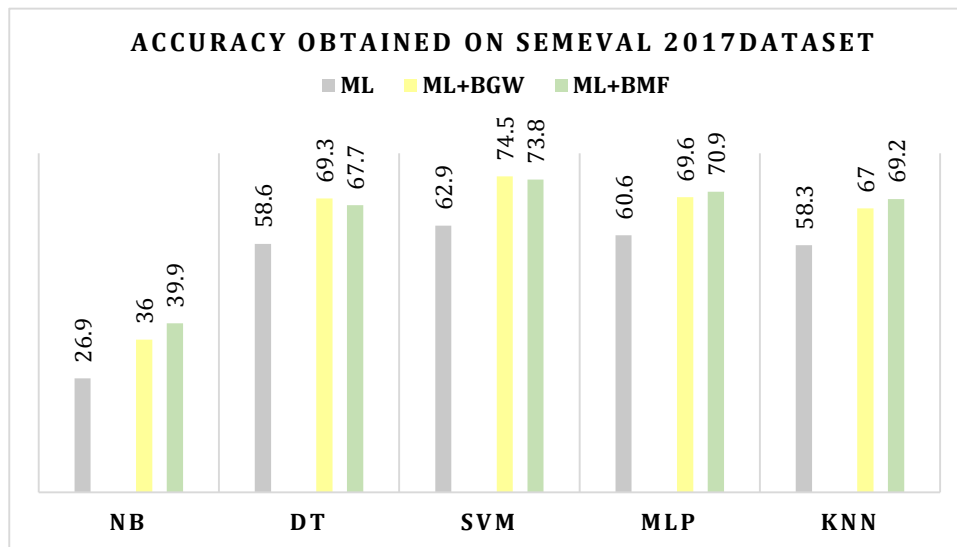
Technique	Non-Optimized Approach (TF-IDF)	Optimized Approach (TF-IDF + BGW)	Features Selected (%)	Optimized Approach (TF-IDF + BMF)	Features Selected (%)
	#Features	#Features	BGW	#Features	BMF
NB	2658	2156	81.1	1411	53.08
DT	2658	1986	74.71	1494	56.20
SVM	2658	2192	82.46	1538	57.86
MLP	2658	1814	68.24	1552	58.38
k-NN	2658	2070	77.87	1447	54.43



**Fig. 4.11.** Features selected using BGW and BMF in SemEval 2017



Figure 4.12 depicts the comparative analysis of application of SI (BGW, BMF) + ML techniques on SemEval 2017 dataset.



**Fig. 4.12.** Accuracy results on using SI+ML on SemEval 2017

From the results, it can be interpreted that optimization algorithms produce more consistent and improved results across datasets when applied for a generic sentiment classification task. The result summary is given in the following table 4.10 for SemEval 2017 dataset.

**Table 4.10:** Result summary for SemEval 2017

Optimizer	Accuracy gain (%)	Features selected (%)
<b>BGW</b>	9.82	76.876
<b>BMF</b>	10.84	58.38

#### 4.3.2.2 Findings of BGW and BMF for SemEval 2016 Dataset

The following table 4.11 depicts the accuracy achieved by baseline classification techniques without and with optimization for the SemEval 2016 dataset.

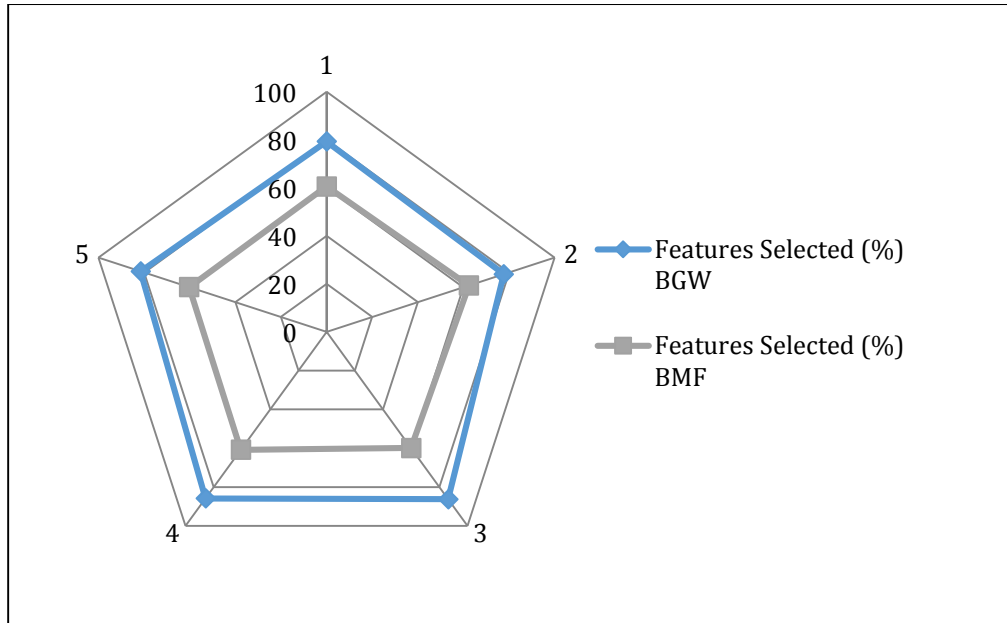
**Table 4.11:** Accuracy gain using BGW and BMF in SemEval 2016

<b>Technique</b>	<b>Non-optimized approach (TF-IDF) Accuracy (%)</b>	<b>Optimized approach (TF-IDF + BGW) Accuracy (%)</b>	<b>Increase in accuracy (%)</b>	<b>Optimized approach (TF-IDF + BMF) Accuracy (%)</b>	<b>Increase in accuracy (%)</b>
<b>NB</b>	29.4	38	8.6	40.3	10.9
<b>DT</b>	54.6	63.9	9.3	64.9	10.3
<b>SVM</b>	65.4	76.5	11.1	74.8	9.4
<b>MLP</b>	59.6	67.9	8.3	70.2	10.6
<b>k-NN</b>	55.3	63.2	7.9	65.9	10.6

The results indicate that the best accuracy with BGW optimization is achieved by SVM on SemEval 2016 dataset, i.e. 76.5%. Amongst all, NB showed the lowest accuracy of around 38% using BGW optimizer. The maximum accuracy gain was obtained by SVM (11.18%) while DT showed an appreciable 9.3% gain in accuracy followed by approximately 8% gain for NB and MLP and k-NN. The average accuracy gain using BGW on SemEval 2016 dataset was of 9.04%.

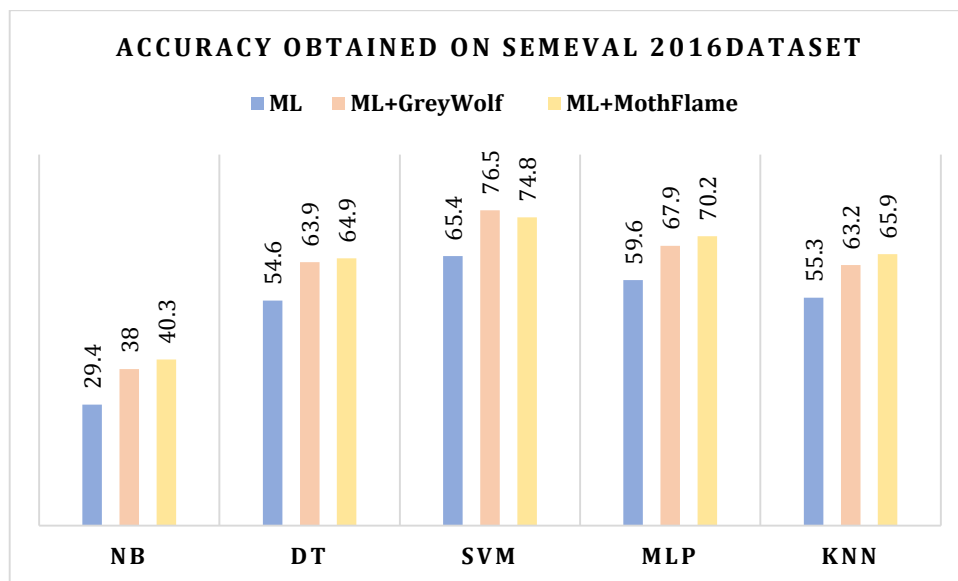
The best accuracy with BMF optimisation is also achieved by SVM on SemEval 2016 dataset, i.e. 74.8%. NB again showed the lowest accuracy of around 40.3% using BMF optimizer. However, the maximum accuracy gain was observed for NB (10.9%) while MLP and k-NN showed an equally appreciable 10.6% gain in accuracy followed by approximately 10% gain for DT and SVM. The average accuracy gain using BMF on SemEval 2016 dataset was of 10.36%.

Table 4.12 displays the number of features selected in SemEval 2016 dataset using the two optimizers. In non-optimised approach all the classification algorithms used the same number of features, which is 2717. After applying BGW for feature selection the minimum number of features selected were 2111 (DT) which is 77.69% selection and maximum were 2375 (SVM) which is 87.41% selection. The analysis shows that on an average 82.13% features were selected. On application of BMF the minimum number of features selected were 1628 (SVM) which is 59.91% selection and the maximum were 1695 (DT) which is 62.38% selection. The results show that on an average 60.79% features were selected, as depicted in figure 4.13.



**Fig. 4.13.** Features selected using BGW and BMF in SemEval 2016

Figure 4.14 depicts the comparative analysis of application of SI (BGW, BMF) + ML techniques on SemEval 2016 dataset.



**Fig. 4.14.** Accuracy results on using SI+ML on SemEval 2016

From the results, it can be interpreted that optimization algorithms produce more consistent and improved results across datasets when applied for a generic sentiment classification task.

The result summary is given in the following table 4.13 for SemEval 2016 dataset.

**Table 4.12:** Result summary for SemEval 2016

<b>Optimizer</b>	<b><i>Accuracy gain (%)</i></b>	<b><i>Features selected (%)</i></b>
<b>BGW</b>	9.04	82.13
<b>BMF</b>	10.36	60.79

#### 4.3.2.3 Findings of PSO for SemEval 2016, 2017 Datasets

The following table 4.13 depicts the accuracy achieved by baseline classification techniques without and with optimization for the SemEval 2016 dataset.

**Table 4.13:** Classifier accuracy before & after feature-selection for SemEval 2016

<b>Classifier</b>	<b>Accuracy before PSO (%)</b>	<b>Accuracy after PSO (%)</b>	<b>Improvement (%)</b>
<b>SGB</b>	59.20	69.43	10.23
<b>RF</b>	58.80	67.64	8.84
<b>ET</b>	58.46	67.03	8.57
<b>GB</b>	58.17	66.43	8.26
<b>Bgg</b>	58.09	65.59	7.50
<b>Bos</b>	57.54	64.87	7.33

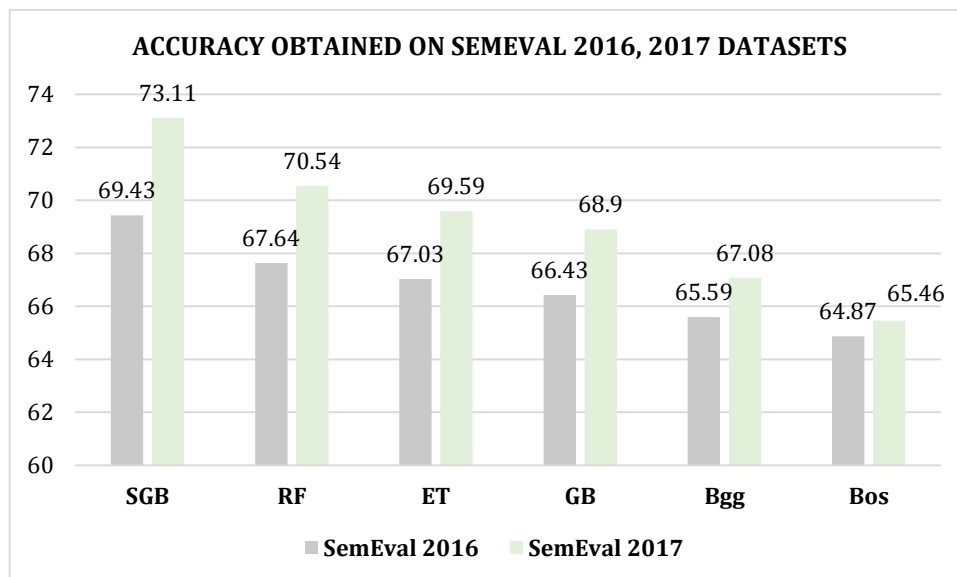
The following table 4.14 depicts the accuracy achieved by baseline classification techniques without and with optimization for the SemEval 2017 dataset.

**Table 4.14:** Classifier accuracy before & after feature-selection for SemEval 2017

<b>Classifier</b>	<b>Accuracy before PSO (%)</b>	<b>Accuracy after PSO (%)</b>	<b>Improvement (%)</b>
<b>SGB</b>	62.60	73.11	10.51
<b>RF</b>	60.91	70.54	9.63
<b>ET</b>	60.85	69.59	8.74
<b>GB</b>	60.41	68.9	8.49
<b>Bgg</b>	59.13	67.08	7.95
<b>Bos</b>	57.96	65.46	7.5

The results of table 4.14 and 4.15 indicate that the maximum improvement in accuracy with optimization is achieved by SGB on both the datasets, i.e. around 10% respectively. Amongst all, Bos showed the minimum improvement in accuracy which is around 7% for both the datasets. After SGB, RF stood next, followed by ET, GB and Bgg. It was observed that the results of improvement in accuracies for Bgg and Bos are comparable for both the datasets. The improvement range after optimization is observed between 7% to 11%.

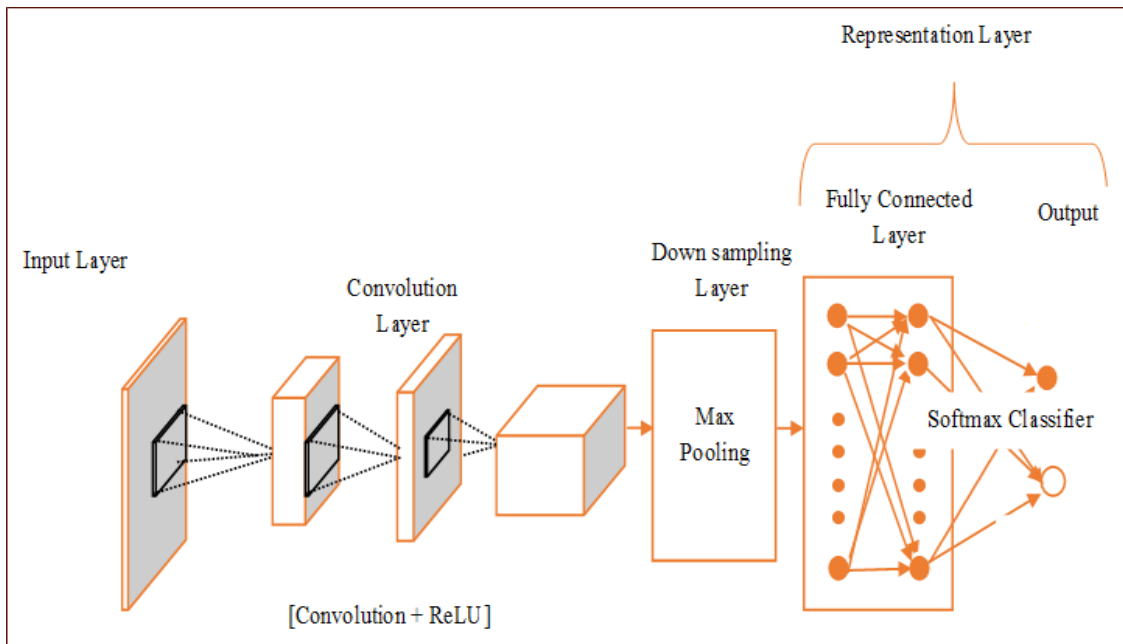
Figure 4.15 depicts the results obtained by application of SI (PSO) + ML (Ensemble) techniques on SemEval 2016, 2017 datasets.



**Fig. 4.15.** Accuracy results on using SI (PSO) + ML on SemEval 2016, 2017

### 4.3.3 Findings of Methodology 3

The SemEval 2016, 2017 datasets were analysed for textual sentiment analysis using CNN (DL). The proposed multi-layer CNN model (as shown in figure 4.16) enhances feature extraction in tweets, which improves the generic sentiment analysis task. Firstly, the tweets from the benchmark dataset are extracted and pre-processed. These pre-processed tweets are connected to the embedding layer, which builds word embeddings using GloVe [102]. The feature representation and extraction in the model is learned in a hierarchical way using word embeddings making it distinctive and better than the lexical or syntactic feature extraction. GloVe is a count-based model of representing words by feature vectors and thus generates a 'word vector table'. It is a log-bilinear model, which studies the relationship of words by counting the number of times they co-occur. Thus, this model aids in mapping all the tokenized words in each tweet to its respective 'word vector table'. The primary aim of embedding layer is to extract relevant features from the input tweets. Proper padding is done for unifying the feature vector matrix. This matrix is given as input to the convolution layer, which consists of multiple filters with a variable size window. The purpose of the layer is to learn high level features from the previous step. The convolution operation is performed to obtain a transformed feature map. ReLU activation function is applied on the feature maps (output of convolution layer) for dealing with the non-linearity in our model. Each filter gives an activation map and this is sent to the next layer, which uses the ReLU activation function. This activation (ReLU) layer introduces non-linearity to the network and generates a Rectified Feature Map. This Rectified Feature Map is fed to the pooling layer to reduce the dimensionality of the feature map. k-max pooling operation is employed which selects the top 'k' features with respect to the various hidden layers and generates a pooled feature map. This pooled feature map is input to the fully connected softmax layer. The output layer is a fully connected layer that consists of softmax activation function. The term Fully Connected implies that each neuron in the previous layer is connected to every single neuron in the next layer. It calculates the probability of any output word thus classifying the tweet sentiment polarity as positive (+1), neutral (0) or negative (-1).



**Fig. 4.16.** Architecture of CNN model

The findings are as follows in table 4.15.

**Table 4.15:** Results using CNN (DL) on SemEval 2016, 2017

Parameters	A	P	R	F
<b>SemEval 2016</b>	85.9	87.8	87.6	87.7
<b>SemEval 2017</b>	86.1	88.3	88.7	88.5

These results of the proposed CNN model were superior and promising when empirically compared with the results of the baseline techniques as shown previous sub-sections. The proposed CNN model had shown accuracy of around 85% to 86% for both the dataset. The recall values for both the datasets was also observed to be above 87%. The model also demonstrated high precision of around 87% to 88% for both SemEval 2016 and 2017 datasets.

It is observed that CNN provides more consistent and improved results across datasets when applied for a generic sentiment classification task. This is primarily because deep CNN does not depend on extensive-manual-feature-engineering [83]. It employs automatic-extensive-feature-extraction mechanism. Earlier conventional models basically relied on hand-crafted-methods-for-feature-extraction based on lexicons etc. which was quite time-consuming and arduous in comparison to dynamic and automated feature selection mechanism adopted by such deep learning method. This aid in learning and

modeling the real world problems more efficiently thus realizing a robust, dynamic and flexible deeper neural architecture. Feature selection is significant in predictive analytics for removing the irrelevant and redundant features, which are not useful. Most real-time datasets such as live message streams suffer from imbalance class distribution (skewed data), where data set is imbalanced. Moreover, Twitter post size characterizes short text with a character-set limit of 280, which is insufficient in providing the desired word co-occurrence data. This data sparseness makes it challenging for the conventional learning models based on statistical features, such as TF-IDF or co-occurrence. The results assert that the use of CNN improves the sentiment classification accuracy. The 'assumption of locality' (autocorrelation) property of CNN can catch the sentiment-classification-information more effectively. It is made possible by retaining the word-order-information and reducing the data sparseness problem. The comprehensibility and generalization of the model makes it easily scalable and applicable to high dimensional real-world problems.

The results clearly assert that the use of CNN improves the sentiment classification accuracy. So, we conclude that deep learning yields enhanced results in comparison to machine learning techniques. Moreover, the accuracy of the machine learning baseline classifiers can be improved using swarm based techniques.

## **4.4 Chapter Summary**

This chapter explicates about sentiment analysis on social media using soft computing techniques. It discusses the overview of the work which entails discussing sentiment analysis predominantly on the scraped and benchmark textual datasets obtained from various social media such as Twitter, Facebook, Tumblr etc. using baseline machine learning techniques. This chapter also focuses on the application of swarm intelligence techniques on the user-generated social media content for enhanced sentiment prediction. It also envisages the application of deep learning technique named CNN on benchmark Twitter dataset for improved sentiment classification.



## **Publications**

- Kumar, A. and Jaiswal, A. Swarm Intelligence Based Optimal Feature Selection for Enhanced Predictive Sentiment Accuracy on Twitter. *Multimedia Tools and Applications, Springer*. 2019 Oct 1; 78(20):29529-53. [**SCIE JOURNAL, Impact factor: 2.313**]. <https://doi.org/10.1007/s11042-019-7278-0>.
- Kumar, A. and Jaiswal, A. Scalable Intelligent Data-Driven Decision Making for Cognitive Cities. *Energy Systems (S.I.: Energy efficiency in building using intelligent computing for smart cities), Springer*. 2019 Nov 19: 1-9. [**ESCI, SCOPUS JOURNAL**]. <https://doi.org/10.1007/s12667-019-00369-5>.
- Kumar, A., Jaiswal, A. Garg, S., Verma, S. Sentiment Analysis Using Cuckoo Search for Optimized Feature Selection on Kaggle Tweets. *International Journal of Information Retrieval Research, (IJIRR)*. 2019 Jan 1; 9(1):1-5. [**ESCI JOURNAL**]. DOI: 10.4018/IJIRR.2019010101.
- Kumar, A. and Jaiswal, A. Deep Network Learning Based Sentiment Classification on User-generated Big Data. *Recent Patents on Computer Science, Bentham Science*, 2019. 12:1. [**SCOPUS JOURNAL**]. <https://doi.org/10.2174/2213275912666190409152308>.
- Kumar, A. and Jaiswal, A. Particle Swarm Optimization-based Ensemble Learning for Enhanced Predictive Sentiment Accuracy of Online-micro Tweets. *In Proceedings of International Conference on Emerging Trends in Information Technology, Springer, 2019*. Pp. 633-646. [**SCOPUS CONFERENCE**]. DOI: 10.1007/978-3-030-30577-2\_56.
- Kumar, A. and Jaiswal, A. Empirical Study of Twitter and Tumblr for sentiment analysis using soft computing techniques. *In Proceedings of the World Congress on Engineering and Computer Science, 2017*. Vol. 1, pp. 1-5. [**SCOPUS CONFERENCE**]. 978-988-14047-5-6.

# Chapter 5

---

*Novel framework for Sentiment Analysis  
using Soft Computing*

# Chapter 5

## Novel framework for Sentiment Analysis using Soft Computing

To compete in the current data-driven economy, it is essential that industrial manufacturers leverage real-time tangible information assets and embrace big data technologies. Data classification is one of the most proverbial analytical technique within the cognitively capable manufacturing industries for finding patterns in structured and unstructured data at the plant, enterprise and industry levels. Motivated by the need to embed analytical capabilities into the core business for real-time value-creation, this paper put forwards a novel cognition-driven social media sentiment mining model. The hybrid model is built on the concord of deep learning (convolution neural network, CNN) and swarm optimized (wolf-search algorithm, WSA) decision tree (DT) for real-time sentiment analytics.

Deep learning techniques have hierarchical learning capabilities but at the same time, the use of adaptive and heuristic optimization to select a near-optimal set of input variables that would minimize variance and maximize generalizability of the learning model, is highly desirable to achieve high prediction accuracy. Motivated by this, we propose the use of a meta-heuristic optimization algorithm, the Wolf Search Algorithm (WSA) for optimal feature selection and build a hybrid model for sentiment classification using deep learning (convolution neural network, CNN) and a swarm (wolf-search algorithm, WSA) optimized decision tree. The proposed deep swarm optimized model embraces the pros of both the techniques where CNN is the automatic feature learner and DT is the sentiment classifier.

The proposed deep swarm optimized model has two primary architectural elements:

- Firstly, a one-dimensional CNN with five layers namely the embedding layer, convolution layer, activation layer, down-sampling layer (pooling layer) and output layer with softmax regression is used to learn distributed feature vector representations of the input.
- Secondly, for final classification, a decision tree classifier is used. This DT takes a combination of the learned vector representations from CNN (the output of top hidden layer) and a meta-heuristically optimized feature vector using WSA to finally output the polarity.

The rationale behind this architecture is that the softmax which is customarily used in CNN to output the probabilities of classes, is a weak classifier that often suffers from difficulty to interpret the results. In classification, predictive probabilities obtained at the end of the pipeline (the softmax output) are often erroneously interpreted as model confidence. It does not express incertitude and may require calibration of predicted probabilities. That is, a model can be uncertain in its predictions even with a high softmax output. Hence the model consists of the following three major components:

- ❖ **CNN for feature learning**
- ❖ **WSA for optimized feature generation**
- ❖ **Feature boosted DT for sentiment classification**

This research presents a cognition driven analytics model, CNN-<sub>WSA</sub>DT, for real-time data classification using three soft computing techniques, namely, deep learning (convolution neural network, CNN), machine learning (decision tree, DT) and swarm intelligence (wolf-search algorithm, WSA). The proposed deep swarm-optimized classifier is a feature boosted decision tree which learns features using a deep convolution net and an optimal feature set built using a meta-heuristic wolf search algorithm. The performance of CNN-<sub>WSA</sub>DT is studied on two benchmark datasets (SemEval 2016, 2017) and the experimental results depict that the proposed cognition model outperforms the other considered algorithms in terms of classification accuracy. Therefore, we seek 5.1. The methodologies and findings for the third research objective is presented in this chapter. A brief summary of above study will ends the chapter.

## 5.1 Research Objective 3

**Research Objective:** To propose a novel framework for sentiment analysis on user-generated content using soft computing techniques.

## 5.2 Methodology

The proposed deep swarm optimized classification model proffers an analytical method that endorses data-driven smart manufacturing.

The architecture of CNN-<sub>WSA</sub>DT model is given in figure 5.1.

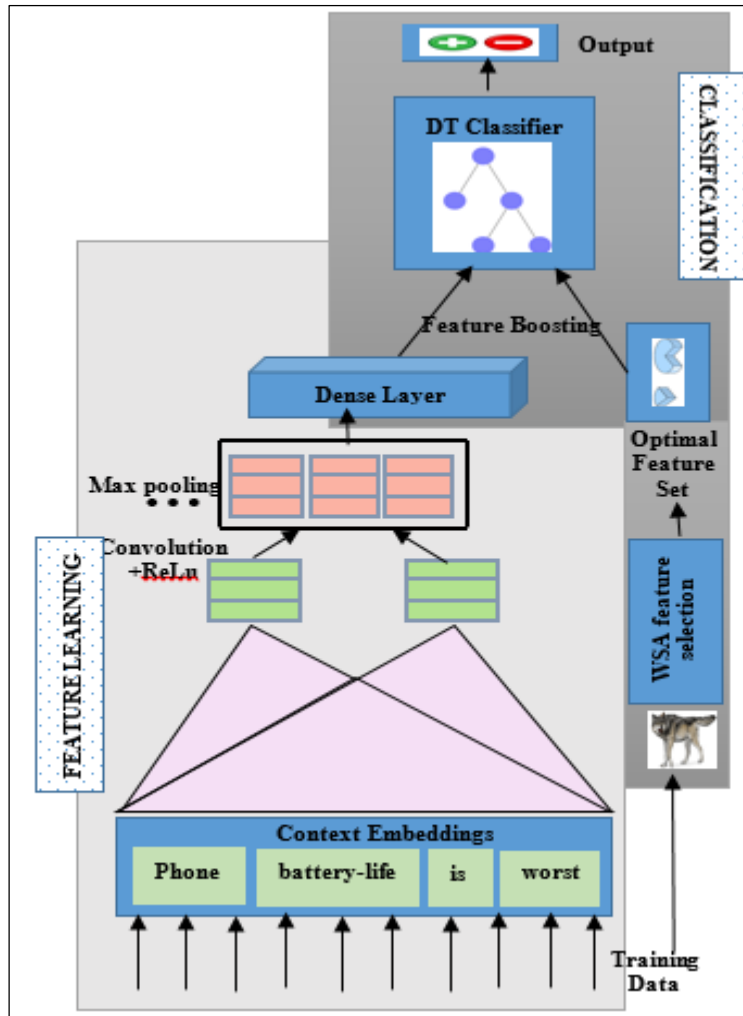


Fig. 5.1. Architecture of proposed CNN-wsaDT model

It consists of the following three architectural components:

- **CNN for feature learning**
- **WSA for optimized feature generation**
- **Feature boosted DT for classification**

The underlying notion that drives this model is that neural architectures are cognitive because they exhibit intelligent behavior by knowing how to categorize, classify, and remember [103-104]. Concurrently, swarms are cognitive systems because they know how to forage, find sites, build nests, and even add and subtract small numbers [105-106].

The first component involves defining, initializing and training of CNN [83]. GloVe [102] is used to generate ‘word vector table’ with an embedding dimension of 300 and a batch size of 50. The model uses three layer convolution architecture with a total of 100 convolution filters each for window size (3, 3). This trains the system to learn the vectors for each word (which would be represented as one hot vector initially) and converts each

word to a vector of integers of 300 dimensions. Textual data is now converted into numerical data for performing convolutions. Padding is used to maintain the fixed input dimensionality feature of CNN, in which zeros are filled in the matrix to get the maximum length amongst all comments in dimensionality. The dropout regularization is set to 0.5 to ensure that that model does not over fit. The default activation function ReLU is applied to the output of convolution layer introducing non-linearity into the model which generates a rectified feature map.

Generic pooling is of varied types such as max, sum, average etc. is used as a down-sampling strategy in convolution networks. In our model, we use max-pooling which selects the top-k features with respect to the multiple hidden layers in order to retain the most significant sentiment feature information. The output derived from the convolution and pooling layers denotes the high-level features of the input tweets. Thus, n-dimensional representation of text is finally obtained which is sent to the output layer for classification. For final classification, a decision tree classifier is used that takes a combination of the learned vector representations from CNN (the output of top hidden layer) and a set of optimal features generated simultaneous by applying WSA on the training data. That is, a boosted feature vector is used to classify the sentiment, thus typifying a deep swarm optimized classification model (as shown below).

**Algorithm 6:** Proposed model (CNN+WSADT)

<b>Hybrid Learning Model (CNN+WSADT)</b>
<p><b>Input:</b> <i>Train, Dev, Test, SemEval</i>- Datasets (2016 &amp; 2017)  <b>Output:</b> <i>Ac</i> – Accuracy obtained</p> <ol style="list-style-type: none"> <li>1: Begin: BuildNet()</li> <li>2: Initialize: InitializeNet(Net)</li> <li>3: Repeat while termination condition is satisfied do</li> <li>4: error <math>\leftarrow</math> TrainNet(Net, Train, Dev)</li> <li>5: End-while</li> <li>6: Select Feature<sub>WSA_opt</sub> <math>\leftarrow</math> WSA(Train, Dev)</li> <li>7: Select Feature<sub>CNN_rel</sub> <math>\leftarrow</math> CNN(Train, Dev)</li> <li>8: Hid<sub>Train</sub> <math>\leftarrow</math> GetTopHiddenLayer(Net, Test)</li> <li>9: Features<sub>concat</sub> <math>\leftarrow</math> Feature<sub>CNN_rel</sub> + Feature<sub>WSA_opt</sub></li> <li>10: Model<sub>DT</sub> <math>\leftarrow</math> DT<sub>Train</sub>(Features<sub>concat</sub>)</li> <li>11: Hid<sub>Test</sub> <math>\leftarrow</math> GetTopHiddenLayer(Net, Test)</li> <li>12: Test<sub>concat</sub> <math>\leftarrow</math> Hid<sub>Test</sub> + Feature<sub>opt</sub></li> <li>13: Ac <math>\leftarrow</math> DT<sub>Test</sub>(Model<sub>DT</sub>, Test<sub>concat</sub>)</li> <li>14: return (Ac)</li> </ol>

Steps 1-5 describe the feature learning using CNN followed by swarm-optimized feature set generation in step 6. Steps 7-13 explicate the feature-boosted classification.

The following sub-sections, present a brief discussion on the techniques used to build the proposed hybrid CNN-wSADT model.

### 5.2.1 Convolution Neural Network (CNN)

CNN is a sequence of convolutional layers, interspersed with activation functions. It is a deep neural architecture which has the power of self-tuning & learning skills by generalizing from the training data. The CNN model enhances feature extraction in tweets, which improves the generic sentiment analysis task [4, 5, 6, 83]. The proposed CNN model comprises of five layers, namely the embedding layer, convolution layer, activation layer, down-sampling layer (pooling layer) and output layer.

The posts from the dataset are pre-processed and input into the embedding layer. The embedding layer of a neural network converts an input from a sparse representation into a distributed or dense representation. In this work, we pre-train the model using GloVe word embedding. The counts matrix is pre-processed by normalizing the counts and log-smoothing them. Thus, this model learn geometrical encodings (vectors) of words in each tweet. Proper padding is done for unifying the feature vector matrix. This matrix is given as input to the convolution layer as shown in figure 5.2.

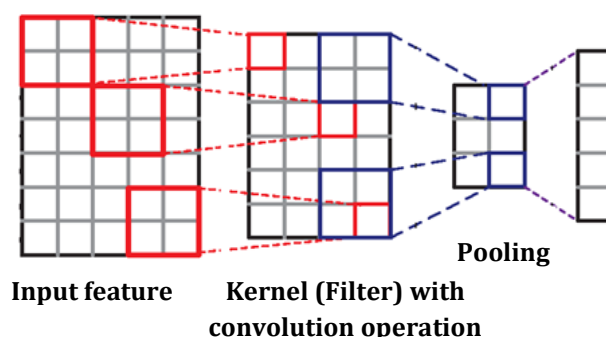


Fig. 5.2.Convolution operation

Every hidden unit consists of three convolution and max pooling layers. The output is therefore a concatenation of the convolution operator over all possible window of words in the tweet. The activation (ReLU) layer is intended to introduce non-linearity to the system and produces a rectified feature map which is inserted into the pooling layer where a max-pooling operation is applied to each convolution. The max-pooling operation extracts the 'k' most important features for each convolution. The output of the final convolution layer, that is, the pooled feature map is a representation of our original input tweet. This representation is then used as an input for decision tree classifier which is combined with the optimized feature set to classify as positive (+1) or negative (-1).

## 5.2.2 Decision Trees (DT)

Decision tree (DT) forms a tree structure to implement the classification or regression models. It continuously splits the data set into smaller subsets based on a criterion to simultaneously generate the tree incrementally. DTs are fast and easy to code, visualize, manipulate, and explain and allow the results to be interpreted very clearly. Other benefits of using decision trees include their application to both numerical and categorical independent variables, efficient handling of missing values in attributes and robustness against skewed distributions.

Softmax regression (logistic regression) is generally used in the fully connected output layer of the CNN but has a single linear boundary unlike DT where we get a non-linear decision boundary. However, when classes are not well-separated, trees are susceptible to over-fitting the training data. Moreover, tree splitting is locally greedy and the DT is more likely to get stuck in local optima. Therefore, to avert being stuck in local optimal we use meta-heuristic optimization.

## 5.2.3 Meta-heuristic Optimization using Wolf Search Algorithm

As one of the key sub-task in data classification, feature engineering is the data manipulation process of using domain knowledge to prepare a compatible dataset for the machine learning algorithm. It includes: feature extraction (n-grams, word2vec, TF-IDF etc.), feature transformation (scaling, median filling etc.) and feature selection (statistical approaches, selection by modelling, grid search and cross validation etc.) [107].

Meta-heuristic optimization algorithms have been progressively studied as wrapper feature selection methods to find candidate solutions in large search spaces. WSA is formulated by simulation of the preying behaviour of wolves. A wolf in WSA hunts independently and rarely joins its peer provided the peer has inhabited a better terrain. WSA can be visualized as multiple individual wolves gathering from various directions towards the optimal solution, instead of a single herd searching for best solution in one direction at a time (figure 5.3).

The natural behaviours of wolves are simulated in WSA as follows [108]:

- The wolves have an unparalleled memory which stores food in caches and track prey. This unparalleled memory is simulated in WSA where each wolf has memory caches that store the positions which are previously visited by it.
- Wolves search for prey during hunting and at the same time they watch out for threats coming towards them. WSA includes a threat probability mechanism which imitates the wolves encounter with enemies. In this condition, the wolf moves away in a random direction by a large distance from its position which prevents getting stuck in local optima.



- Wolves have an outstanding judgment of smell which helps them to locate prey. WSA simulates this by enabling each wolf to have a sensing distance that creates a coverage area which is called visual distance. While searching when a wolf is not able to find food (the global optimum) or a better terrain than its current position within visual range, the wolves move in Brownian motion.

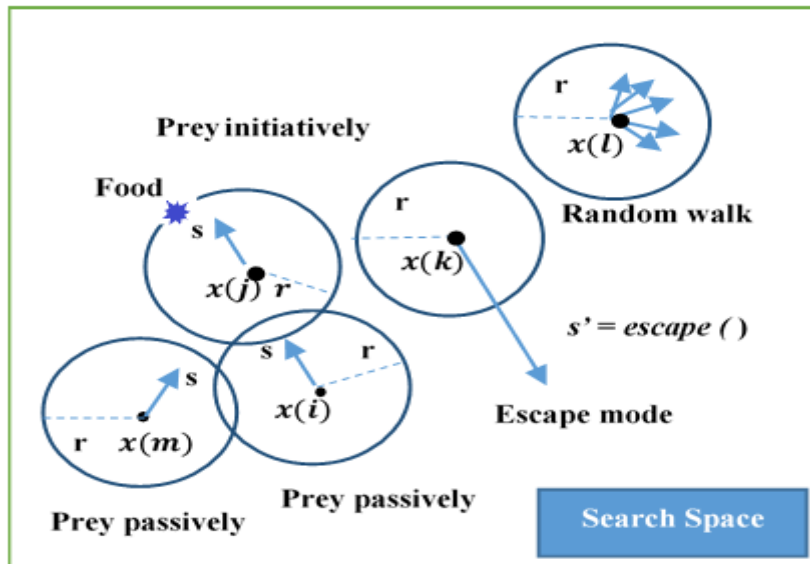


Fig. 5.3.WSA in action

Algorithm for WSA is shown below:

#### Algorithm 7: WSA

WSA
<p><b>Objective function</b> <math>f(x)</math>, <math>x = (x_1, x_2, x_3, \dots, x_d)^T</math></p> <p><b>Initialize</b> the population of wolves, <math>x_i</math> (<math>i = 1, 2, 3, \dots, W</math>)</p> <p>Define and initialize parameters:</p> <p><math>r</math> = radius of the visual range</p> <p><math>s</math> = step size by which a wolf moves at a time</p> <p><math>\alpha</math> = velocity factor of wolf</p> <p><math>p_a</math> a user-defined threshold <math>[0 \dots 1]</math>, determines how frequently an enemy appears</p> <p>While (<math>t &lt; \text{generations}</math> &amp;&amp; stopping criteria not met)</p> <p>For <math>I=1: W</math> // for each wolf</p> <p><i>Prey_new_food_initiatively</i> ();</p> <p><i>Generate_new_location</i> (); // check whether the next location suggested by the random number generator is new.</p> <p>If not, repeat generating random location</p> <p>If <math>\text{dist}(x_i, x_j) &lt; r</math> &amp;&amp; <math>x_j</math> is better as <math>f(x_i) &lt; f(x_j)</math></p>

```

xi moves towards xj // xj is a better than xi
Else-if
xi = Prey_new_food_passively ();
End-if
Generate_new_location ();
If rand () > pa
xi = xi + rand() + v; // escape to a new position
End-if
End-for
End-while

```

The WSA follows some rules which are given below:

- i. The wolves have a visual distance with a radius as  $v$  and  $X$  as a set of continuous possible solutions. In hyper-plane, this distance would be estimated by Minkowski distance as given in (5.1):

$$v \leq d(x_i, x_c) = \left( \sum_{k=1}^n |x_{i,k} - x_{c,k}|^\lambda \right)^{1/\lambda}, x_c \in X \quad (5.1)$$

where  $x_i$  is the current position,  
 $x_c$  are all the potential neighboring positions near  $x_i$ ,  
 $\lambda$  is the order of the hyper space.

- ii. The quality of a wolf's current position is given by the fitness of the objective function. The wolf continually attempts to relocate to better terrain inhabited by a companion and will finally choose best terrain in case of multiple better terrains. Else, the wolf will continue to move randomly in Brownian motion.
- iii. When a wolf senses an enemy, it will escape to a random position beyond its visual range to move away from the threat.

### 5.2.3.1 Merging with Other Wolves

In WSA the fitness of the objective function determine the quality of wolf's current position. A wolf always wants to be in a position where there is greater probability of finding a prey (food) and lower probability of meeting a predator (being hunted) and it will rarely move into territory occupied by another wolf if that territory is better. This works as follows:

Initially, each wolf locates other wolves within its visual range and evaluates the quality of position of each of its companions. The best location amongst all is compared

with the wolf's position. If it is beneficial to locate to this new position, the wolf relocates and prey there. Otherwise, the wolf searches in a Brownian motion with an incremental step size.

The implementation of this movement is shown in (equation 5.2):

$$x(i) = x(i) + \beta_o e^{-r^2} (x(j) - x(i)) + \text{escape}() \quad (5.2)$$

where,

$\text{escape}()$  generates a random position that enables the wolf to hop,

$x(i)$  is the location of the wolf,

$x(j)$  is the neighbor that is in a better position,

$\beta_o e^{-r^2}$  is the incentive formula which represents the betterment (gain) achieved by wolf by moving to a new position, where  $\beta_o$  is the origin of food,  $r$  is the distance between the wolf and the new position.

If there are no better terrains occupied by wolf's peers and the wolf is only in the best position, the other wolves will ultimately crowd to the wolf's current position.

### 5.2.3.2 Preying

Typically, a wolf looks out a region completely to search for food in a pattern of Brownian motion.

WSA exhibits three different kinds of preying behaviour:

- i. **Preying initiatively:** The objective of the optimization function is represented as food. In this step each wolf checks its visual range to detect prey. The wolf will move step by step in the direction of the prey detected with highest fitness.
- ii. **Prey passively:** in case the wolf is not able to find food or better position occupied by a peer in the preceding step, it will prey passively by staying alert for incoming threats and also it will check the position of its peers in an attempt to improve its current position.
- iii. **Escape:** the wolf escapes quickly when a threat is detected. It relocates itself to a random new position such that its escape distance is greater than its visual range. Escape prevents all the wolves from getting stuck at a local optimum.

These preying steps can be defined mathematically given in (equation 5.3):

$$\text{if moving} = \begin{cases} x(i) = x(i) + \alpha.r.\text{rand}() & //\text{Prey} \\ x(i) = x(i) + \alpha.s.\text{escape}() & //\text{Escape} \end{cases} \quad (5.3)$$

where  $x(i)$  is the position of the wolf,

$\alpha$  is the velocity,

$\text{rand}()$  is a random function with mean value in  $[-1,1]$ ,

$v$  is the visual distance,

$s$  is the step size,

$escape()$  is a custom function that generates a position in a random manner which is greater than  $v$  and less than half of the solution boundary.

The parameters for WSA were set as: population size = 20; iterations=20; chaotic co-efficient=4.

## 5.3 Findings

The proposed CNN- $w_{SA}DT$  was evaluated for classification performance accuracy (Ac) (in percentage). Two benchmark Twitter datasets, SemEval 2016 (Task 4, subtask-A) and SemEval 2017 (Task 4, subtask-A) were used for training and validation. The tweets were labelled as positive, negative, and neutral. Both SemEval 2016 & 2017 are unbalanced datasets with SemEval 2017 dataset comprising of 2352 positive, 3811 negative, and 5742 neutral tweets, and SemEval 2016 dataset consisting of 7059 positive, 3231 negative and 10341 neutral tweets. The classification results were assessed by partitioning the dataset into training and test sets. 10-fold cross-validation was performed to create a validation set and find the best parameters. Scikit-learn library and Keras deep learning library with Theano backend were used for implementation.

### 5.3.1 Performance of the proposed CNN + $w_{SA}DT$

The accuracy reported by the hybrid model was approximately 90% for both the datasets. This is primarily because CNN does not depend on extensive manual feature engineering. It employs automatic extensive feature extraction mechanism. This aids in learning and modelling the real-world problems more efficiently thus realizing a robust, dynamic and flexible deeper neural architecture. Also, the application of WSA optimization produced a set of optimized feature which were combined with the pooled feature of CNN to train the DT classifier for improved classification accuracy. Table 5.1 depicts the accuracy results achieved.

**Table 5.1:** Results of CNN +  $w_{SA}DT$

Data Set	Accuracy (Ac)
SemEval 2016	89.4%
SemEval 2017	89.7%

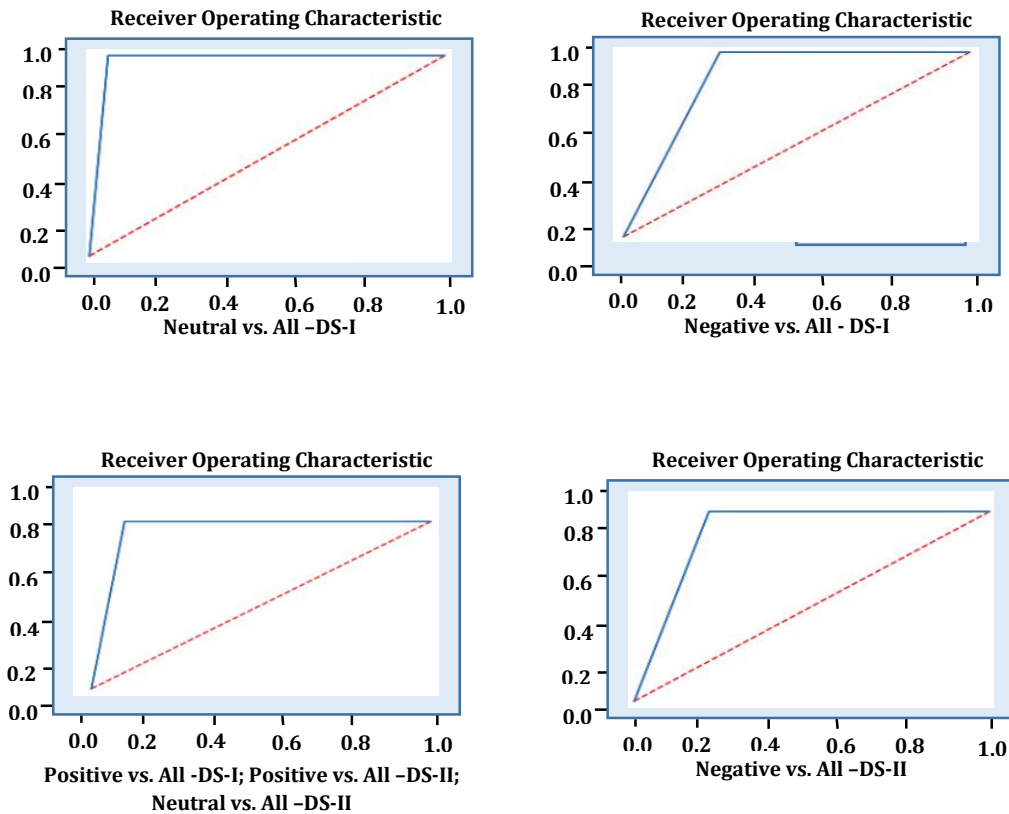
To highlight the improvement shown by the proposed model, we evaluated the CNN model independently as baseline on the both the data sets.

The proposed CNN+ $w_{SA}DT$  model achieves nearly 3.5% more prediction accuracy, as shown in table 5.2.

**Table 5.2:** Comparative analysis of CNN and hybrid model for SemEval 2016, 2017

Dataset	CNN	CNN+ <sub>WSA</sub> DT
SemEval 2016	85.9	89.4
SemEval 2017	86.1	89.7

The hybrid model is built on the concord of deep learning (convolution neural network, CNN) and swarm- optimized (wolf-search algorithm, WSA) decision tree (DT) for real-time sentiment analytics. Based on the results obtained by application of aforesaid soft computing techniques, in our research, we observed that our proposed hybrid model DL+SI+ML using CNN, wolf search algorithm and decision tree, produced enhanced results when compared to all the above. Figure 5.4 depicts the AUC-ROC curves for DS-I (SemEval 2016) & DS-II (SemEval 2017).

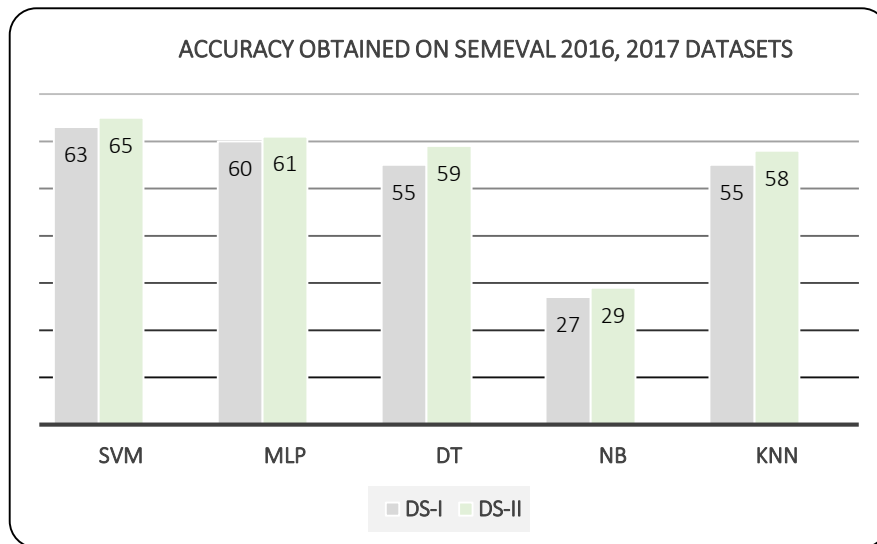


**Fig. 5.4.** ROC for SemEval 2016 (DS-I) and SemEval 2017 (DS-II)

### 5.3.2 Comparison of DT with other ML techniques

To endorse the use of DT, it was compared with four other supervised machine learning techniques, namely, support vector machines (SVM), naïve bayes (NB), k-nearest neighbour (kNN), and multi-layer perceptron (MLP) on both the datasets. Term frequency-inverse document frequency (TF-IDF) weighting [87] was used to construct

the features set used to train the classifiers. SVM achieved the highest accuracy with 63% for DS-I (SemEval 2016) and 65% for DS-II (SemEval 2017). MLP also depicted encouraging results with an accuracy of around 60% and 61% for DS-I (SemEval 2016) and DS-II (SemEval 2017) respectively. Next to MLP, DT attained an accuracy of 55% and 59% for DS-I (SemEval 2016) and DS-II (SemEval 2017) respectively. Figure 5.5 depicts the comparative analysis of the aforesaid techniques based on accuracy (percentage).



**Fig. 5.5.** Comparison of supervised learning techniques using accuracy

Although, SVM individually showed the highest accuracy, for the proposed model, we preferred choosing DT so as to develop a robust model for sentiment analysis which attune to the ‘skewness’ in real-time datasets. Also, while MLP came next to SVM, but as it is neural model and our model is already using a deep layered neural architecture, the CNN, we opted using DT for final classification in our proposed hybrid model.

Experiments were also done to discern the selection of optimal subset of features using WSA with these supervised learning techniques. Table 5.3 depicts the number and percentage of features selected in both the datasets using the five different classifiers and WSA optimization.

**Table 5.3:** Feature selection using TF-IDF+WSA

Techniques	DS-I (TF-IDF)	(TF-IDF + WSA)	%	DS-II (TF-IDF)	(TF-IDF + WSA)	%
<b>NB</b>	2717	2156	79.35	2658	2156	81.10
<b>DT</b>	2717	2095	77.10	2658	1986	74.71
<b>SVM</b>	2717	2375	87.41	2658	2192	82.46
<b>MLP</b>	2717	2332	85.82	2658	1814	68.24
<b>k-NN</b>	2717	2215	81.52	2658	2070	77.87

Quite clearly, our choice was upheld to DT which reduced the search space notably by integrating its linearly separable advantage to the non-linear search capability of WSA. The average feature selection by WSA optimized DT was approximately 76% for both the datasets.

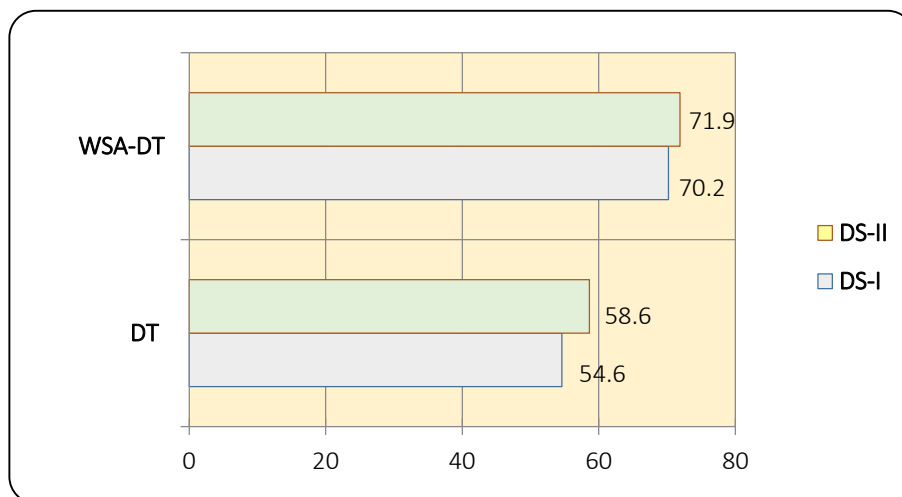
### 5.3.2 Comparison of WSA with other meta-heuristic optimization algorithms

We conducted an empirical analysis to validate the use of meta-heuristic WSA. That is, five optimization algorithms, namely binary grey wolf (BGW), binary moth flame (BMF), and wolf search algorithm (WSA) were used to generate the optimal feature subset and the DT was trained using the optimal feature set to witness the viable classification accuracy. Table 5.4 shows the comparison of accuracy achieved for each of the optimization algorithm used to train DT and it is observed that WSA optimized DT outperforms the others.

**Table 5.4:** Accuracy comparison of optimization techniques

Datasets	BGWDT	BMFDT	WSADT
SemEval 2016	63.9	64.9	70.2
SemEval 2017	69.3	67.7	71.9

The graph in figure 5.6 shows the accuracy of DT in percentage with and without WSA optimization.



**Fig. 5.6.** Accuracy (%) of DT with and without WSA optimization

Thus, an average accuracy gain of 14.5% and an average feature reduction of 24% was observed for DT with WSA optimization.

## 5.4 Chapter Summary

This chapter specifically aims at developing a novel hybrid model for real-time sentiment classification harnessing the best of three diverse domains of soft computing, namely, the deep learning (DL), machine learning (ML) and the swarm intelligence (SI). A novel model for enhanced sentiment prediction using SI+DL+ML techniques is also discussed in detail.

### **Publications**

- Kumar, A. and Jaiswal, A. A Deep Swarm Optimized Model for leveraging Industrial Data Analytics in Cognitive Manufacturing. *IEEE Transactions on Industrial Informatics*. 2020 Jun 29. [SCIE JOURNAL, Impact factor: 9.112]. DOI: 10.1109/TII.2020.3005532.



# Chapter 6

---

---

## *Conclusion and Future Scope*

# Chapter 6

## Conclusion and Future Scope


This chapter recaps the research summary. It also highlights the mapping of the research objectives with list of publications. A thorough discussion of limitations of the study and future scope has been discussed. And finally, conclusion concludes the thesis.

### 6.1 Research Summary

Classifying sentiments in online social data is a typical natural language processing problem which takes text as input and converts these inputted texts into features that the learning algorithms can understand. Selecting features is one of the most *difficult* and imprecise part of the machine learning. The features in given data are important to the predictive models used and influence the results achieved. The quality and quantity of the features have great influence on whether the model is good or not. Though results achieved also depend on the model and the data and not just the chosen features, nevertheless choosing the right features is still very important. Better features can produce simpler and more flexible models, and they often yield better results. Given the infinite number of potential features, it's often not computationally feasible for even the most sophisticated algorithms. The objective of this research is thus to process the valuable, hidden information from raw, uncertain, imprecise and high-dimensional social media data into a form more amenable to learning and maximizing predictive power using soft computing techniques. It specifically aims at developing a novel hybrid model for real-time sentiment classification harnessing the best of three diverse domains of soft computing, namely, the deep learning (CNN), machine learning (DT) and the swarm intelligence (SI).

Table 6.1 depicts the mapping of research objectives (RO) with the list of publications:

**Table 6.1:** Mapping of research objectives with the list of publications

RO	RO detail	Published Paper
I	To seek the convergence of Web 2.0 technologies and sentiment analysis on social media for real-life applications.	 Kumar, A. and Jaiswal, A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. <i>Concurrency and Computation: Practice and Experience</i> , Wiley. 2020 Jan 10:32(1):e5107. [SCIE JOURNAL, Impact factor: 1.447]. <a href="https://doi.org/10.1002/cpe.5107">https://doi.org/10.1002/cpe.5107</a> .

<p><b>II</b></p>	<p>To perform sentiment analysis on textual unstructured data on the Web.</p>	<ul style="list-style-type: none"> <li data-bbox="659 192 1355 465">✚ Kumar, A. and Jaiswal, A. Swarm Intelligence Based Optimal Feature Selection for Enhanced Predictive Sentiment Accuracy on Twitter. <i>Multimedia Tools and Applications</i>, Springer. 2019 Oct 1; 78(20):29529-53. [SCIE JOURNAL, Impact factor: 2.313]. <a href="https://doi.org/10.1007/s11042-019-7278-0">https://doi.org/10.1007/s11042-019-7278-0</a>.</li> <li data-bbox="659 510 1355 784">✚ Kumar, A. and Jaiswal, A. Scalable Intelligent Data-Driven Decision Making for Cognitive Cities. <i>Energy Systems (S.I.: Energy efficiency in building using intelligent computing for smart cities)</i>, Springer. 2019 Nov 19: 1-9. [SCOPUS JOURNAL, Impact factor: 1.65]. <a href="https://doi.org/10.1007/s12667-019-00369-5">https://doi.org/10.1007/s12667-019-00369-5</a>.</li> <li data-bbox="659 828 1355 1052">✚ Kumar, A. and Jaiswal, A. Deep Network Learning Based Sentiment Classification on User-generated Big Data. <i>Recent Patents on Computer Science</i>, Bentham Science, 2019. 12:1. [SCOPUS JOURNAL]. <a href="https://doi.org/10.2174/2213275912666190409152308">https://doi.org/10.2174/2213275912666190409152308</a>.</li> <li data-bbox="659 1097 1355 1321">✚ Kumar, A. and Jaiswal, A. Empirical Study of Twitter and Tumblr for sentiment analysis using soft computing techniques. In <i>Proceedings of the World Congress on Engineering and Computer Science</i>, 2017. Vol. 1, pp. 1-5. [SCOPUS CONFERENCE]. 978-988-14047-5-6.</li> <li data-bbox="659 1366 1355 1680">✚ Kumar, A. and Jaiswal, A. Particle Swarm Optimization-based Ensemble Learning for Enhanced Predictive Sentiment Accuracy of Online-micro Tweets. In <i>Proceedings of International Conference on Emerging Trends in Information Technology</i>, Springer, 2019. Pp. 633-646. [SCOPUS CONFERENCE]. DOI: 10.1007/978-3-030-30577-2_56.</li> </ul>
<p><b>III</b></p>	<p>To propose a novel framework for sentiment analysis on user-generated content using soft computing techniques.</p>	<ul style="list-style-type: none"> <li data-bbox="659 1727 1355 1951">✚ Kumar, A. and Jaiswal, A. A Deep Swarm Optimized Model for leveraging Industrial Data Analytics in Cognitive Manufacturing. <i>IEEE Transactions on Industrial Informatics</i>. 2020 Jun 29. [SCIE JOURNAL, Impact factor: 9.112]. DOI: 10.1109/TII.2020.3005532.</li> </ul>

## 6.2 Limitations of the study

With the growing use of customary memes and GIFs in the social feeds, different modalities and the heterogeneity of the multi-modal data poses new challenges for sentiment analysis. Also, the cultural diversities, country-specific trending topics and hash-tags on social media and easy availability of native language keyboards for social media applications add to the variety and volume of user-generated content. Thus, as promising future direction, models and benchmark datasets for multi-lingual, multi-modal sentiment analysis research task are ardently desired.

## 6.3 Future Scope

The entire process of text classification can be enhanced by improving the task of opinion classification with the aid of exploiting other soft computing techniques and by examining and substantiating various combinations of hybrid classifiers. Intelligent adaptive models are required to deal with the information overload on the chaotic and complex social media portals. Most of the real-time datasets such as tweets suffer from imbalance class distribution (skewed data), it encourages to investigate the use of deep learning techniques for learning data representations. The use of mash-up languages and novelty in vocabulary add to the challenges of prediction of opinion and characterize some open problems for future research within the domain. Most of the work done to analyse the sentiments is on textual un-structured web data, whereas multimedia (non-textual: audio, video, image, gif's, emoticons) sentiment analysis has not been explored much. Fine-grain sentiment analysis which includes, emotion analysis, sarcasm detection, rumour detection, irony detection have been identified as potential directions of research. Social media has become an informal way of communication with increased use of slangs and emoticons, mal-formed words, colloquial expressions, thus increasing the dimensionality, fuzziness and complexity of the content. As promising future direction, the proposed model (CNN + wsADT) can be used to implement the classification algorithm in the programming model like Map Reduce to achieve parallel processing, thereby solving the problems of hardware and communication overhead for managing large-scale and streaming datasets. It could also be used for gauging the user-generated big-data that will allow learning context and facilitate cognitive design for mass personalization as well. Also, as the model adds a layer of interpretability, its prospects as an explainable AI solution needs further discussion.

## 6.4 Conclusion

Sentiment analysis is one of the most proverbial and easily understandable applications of user-generated big data facilitating decision making. Analysing and classifying sentiments from Web data is challenging as effective feature selection is computationally

hard. The results of our study suggested that the best sentiment accuracy and precision was achieved using support vector machine (SVM) for both Twitter (tweets) and Tumblr (tumblogs). SVM outperformed all other supervised classification algorithms in terms of precision and accuracy, followed by naïve bayesian (NB) and k-nearest neighbour (k-NN) techniques, with the highest precision for SVM, (approximately 76% for Twitter and approximately 78% for Tumblr). Multi-layer perceptron neural network (NN) also showed encouraging results, quite akin to NB for both the micro-blogs. Next to NN, decision trees (DT) had shown comparable accuracy. The observed variation in the results was merely due to the fact that large chunks of tweets and tumblogs were skewed towards negative sentiments, predominantly for the topics like Rio Olympics, US presidential elections and Donald Trump's claim for Muslim ban. From the results it was deduced that improved and optimized results were observed for Tumblr in contrast to Twitter.

Also, we observed that Support Vector Machines (SVM) had attained the highest accuracy in the range of 63% to 65% whereas naïve bayesian (NB) reported the lowest accuracy of around 50% to 55% when applied to SemEval 2016, 2017 datasets. It was also deduced that the multilayer perceptron (MLP) showed encouraging results by producing accuracy of around 60% to 61% for SemEval 2016, 2017 datasets respectively. Next to MLP, came decision tree (DT) that showed accuracy within the range of 55% to 59% for SemEval 2016, 2017 datasets. After this, k-nearest neighbor (kNN) followed which yielded accuracy of approximate 55% and 58% for SemEval 2016, 2017 datasets.

When ensemble techniques were used, the pragmatic results depicted that the maximum A was attained by SGB on both the datasets (SemEval 2016, 2017), i.e. around 59% and 63% respectively. Amongst all, Bos showed the minimum accuracy of around 57% for both the datasets. After SGB, RF showed an appreciable classification accuracy of around 58% for SemEval 2016 and 61% for SemEval 2017, followed by ET, GB and Bgg. It was observed that the results of ET and GB were comparable for both the datasets.

Furthermore, when IMDb movie reviews and Sentiment140 datasets were used for analysis, it was observed that amongst all movie reviews, majority of the reviews were towards higher range of viewer satisfaction i.e. within the range of approximately 68% to 82%. This higher level showed that these reviews were correct and absolute that could further be utilized by the viewers for judicious decision making, whereas there were around 32% to 18% reviews that happened to be quite inaccurate in sentiment prediction. It all depended on the technique employed as different techniques yield different results based on their computational abilities. Amongst all, LogR predicted sentiments most aptly, i.e. it was able to accurately predict 82% of the sentiments as positive, negative or neutral and for rest of the cases, it failed to predict absolutely. Likewise, for the other dataset (sentiment140), different techniques produced different results for sentiment prediction about the opinions for any topics, brand etc. Highest accuracy of around 91% was obtained using support vector machines (SVM) that signifies

that for 91% of the times, it was able to correctly classify the sentiment as either positive or negative about any brand or topic etc., and for rest of the cases, it went out inaccurate. Consequently, we could conclude that different techniques showed different sentiment prediction rate owing to their respective competence and ability. The techniques that were showing higher accuracy percentages offer better credibility to the users.

When swarm intelligent techniques were used for feature selection optimization, it was observed that the best accuracy was reported by binary grey wolf (BGW) optimization for SVM, i.e. 74.5% and NB showed the lowest accuracy of around 36%. The maximum accuracy gain was obtained by SVM (11.6%) while DT showed an appreciable 10.7% gain in accuracy followed by approximately 9% gain for NB and MLP and k-NN. The average accuracy gain using BGW on SemEval 2017 dataset was of 9.82%. For the other dataset, the results also specified that the best accuracy with BGW optimization was achieved by SVM (on SemEval 2016 dataset), i.e. 76.5%. Amongst all, NB showed the lowest accuracy of around 38% using BGW optimizer. The maximum accuracy gain was obtained by SVM (11.18%) while DT showed an appreciable 9.3% gain in accuracy followed by approximately 8% gain for NB and MLP and k-NN. The average accuracy gain using BGW on SemEval 2016 dataset was of 9.04%.

When binary moth flame (BMF) was applied, the results indicated that the best accuracy with optimisation was again achieved by SVM on SemEval 2017 dataset, i.e. 73.8%. NB showed the lowest accuracy of around 39.9% using BMF optimizer. The maximum accuracy gain was observed for NB (13%) while SVM and k-NN showed an equally appreciable 10.9% gain in accuracy followed by approximately 10% gain for DT and MLP. The average accuracy gain using BMF on SemEval 2017 dataset was of 10.84%. The best accuracy with BMF optimisation was also achieved by SVM on SemEval 2016 dataset, i.e. 74.8%. NB again showed the lowest accuracy of around 40.3% using BMF optimizer. However, the maximum accuracy gain was observed for NB (10.9%) while MLP and k-NN showed an equally appreciable 10.6% gain in accuracy followed by approximately 10% gain for DT and SVM. The average accuracy gain using BMF on SemEval 2016 dataset was of 10.36%.

When particle swarm optimization (PSO) was applied for optimization using ensemble methods, the results reported that the maximum improvement in accuracy with optimization was achieved by SGB on both the datasets (SemEval 2016, 2017), i.e. around 10% respectively. Amongst all, Bos showed the minimum improvement in accuracy which was around 7% for both the datasets. After SGB, RF stood next, followed by ET, GB and Bgg. It was observed that the results of improvement in accuracies for Bgg and Bos are comparable for both the datasets. The improvement range after optimization was observed to be in between 7% to 11%. From these results obtained via application of SI (optimization) techniques, we can conclude that optimization algorithms produced more consistent and improved results across datasets when applied for a generic sentiment classification task.

Moving over to application of DL to datasets, the results showed that the applied CNN model was superior and promising when empirically compared with the results of the baseline techniques as shown in previous sub-sections. The proposed CNN model had shown accuracy of around 85% to 86% for both the dataset (SemEval 2016, 2017). The recall values for both the datasets was also observed to be above 87%. The model also demonstrated high precision of around 87% to 88% for both SemEval 2016 and 2017 datasets. Thus, we conclude that deep learning yields enhanced results in comparison to Machine Learning methods for sentiment prediction. Also, it was observed that the accuracy of the machine learning baseline classifiers can be improved using swarm based techniques.

Now, coming to our proposed model CNN+<sub>WSA</sub>DT, it was observed that it achieved nearly 3.5% more prediction accuracy in comparison to above applied techniques. This research proffered a novel hybrid model for real-time sentiment classification harnessing the best of three diverse domains of soft computing, namely, the deep learning (CNN), machine learning (DT) and the swarm intelligence (SI). The architectural components of our proposed model comprised of CNN with five layers namely the embedding layer, convolution layer, activation layer, down-sampling layer (pooling layer) and output layer with softmax regression. And for the final classification, we employed a decision tree (DT) classifier. This DT takes a combination of the learned vector representations from CNN and a meta-heuristically optimized feature vector (obtained via application of WSA) to finally output the sentiment polarity. Although, the results of our research exhibited that SVM individually showed the highest accuracy (among all other applied baseline supervised methods), still, for the proposed model, we preferred choosing DT. It was mainly because of the fact that we intend to develop a robust model for sentiment analysis which could attune itself to the 'skewness' associated with the real-time datasets. The rationale behind this architecture is that the softmax regression (logistic regression) which is customarily used in CNN to output the probabilities of classes, is a weak classifier that often suffers from difficulty to interpret the results. Moreover, no weightage is given to relevant features as simply word embeddings are used for all features. Thus, in the proposed CNN-WSADT model, the decision tree is trained using a boosted feature vector obtained by combining the CNN trained features and WSA optimized feature vector. In our study, we opted to replace the softmax layer with decision tree as primarily it makes the model easy to interpret. Additionally, decision tree splits the input space into hyper-rectangles according to the target, it does not suffer from imbalanced support vector ratio or soft margin optimization problems which are commonly observed in classifiers like support vector machines (SVM). But on the flip side, DT has a likelihood of reaching a locally optimal solution as it is a top-down algorithm with a divide and conquer approach. Over-fitting of the training data can negatively affect the modelling power of the technique and relegate the predictive accuracy. Population-based meta-heuristics, especially the ones inspired by nature have helped solving different optimization problems and been used successfully for feature selection in many applications. Therefore, for the proposed

model, we primarily intend to generate the optimal feature set. We firstly used the conventional TF-IDF (term frequency-inverse document frequency) feature extraction and then used a meta-heuristic optimization algorithm, the wolf search algorithm (WSA) to select the most relevant set of features. WSA imitated the way wolves searched for food, survived and avoided enemies. WSA possessed individual local searching ability and autonomous flocking movement in tandem. That is, each wolf is an independent hunter with its own behaviour and only joins the peer when peer is in superior place within its visual range. The hypothesis behind WSA is: rather than looking for the best solution in one direction by forming a single pack/herd, it considered many leaders swarming to the optimal solution from several directions. Also, to avert trapping in local optima, the appearance of a hunter (threat/enemy) corresponding to each wolf is randomly added such that the wolf escapes from the hunter's visual range to strive for better solutions within in the search space. Also, we could observe that majorly MLP came next to SVM in terms of sentiment classification accuracy for real-time datasets, but as it (MLP) is neural model and our model is already using a deep layered neural architecture, the CNN, we opted using DT for final classification in our proposed hybrid model. Experiments were also done to discern the selection of optimal subset of features using WSA with these supervised learning techniques. Quite clearly, our choice was upheld to DT which reduced the search space notably by integrating its linearly separable advantage to the non-linear search capability of WSA. The average feature selection by WSA optimized DT was approximately 76% for both the datasets. The proposed deep swarm optimized classification model proffers an analytical method that endorses data-driven smart manufacturing. The underlying notion that drives this model is that neural architectures are cognitive because they exhibit intelligent behaviour by knowing how to categorize, classify, and remember. Concurrently, swarms are cognitive systems because they know how to forage, find sites, build nests, and even add and subtract small numbers. Based on this, a novel cognition-driven data classification model was put forward which embeds predictive analytics capabilities into the core manufacturing task for real-time value-creation. This proposed hybrid model for real-time sentiment classification uses CNN and wolf-search optimized decision tree trained and validated on two benchmark Twitter datasets. The combined optimal feature vector generated a superior learning model with an average accuracy of 89.5% validated on both datasets. Ultimately, gauging this user-generated big-data will allow learning context and facilitate cognitive design for mass personalization.



## REFERENCES

1. A. Kumar, K. Srinivasan, W.H. Cheng, & A.Y. Zomaya, "Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data". *Information Processing & Management*, Elsevier, vol. 57, no.1, 102141, 2020.
2. C.C. Aggarwal, C. Zhai, *A Survey of Text Classification Algorithms*, in: *Min. Text Data*, 2012.
3. M. Rajman, R. Besançon, *Text Mining - Knowledge extraction from unstructured textual data*, Proc. 6th Conf. Int. Fed. Classif. Soc. (1998) 473–480.
4. A. Kumar, T.M. Sebastian, *Sentiment analysis on Twitter*, *IJCSI Int. J. Comput. Sci. Issues*. 9 (2012) 372–378.
5. B. Pang, L. Lee, *Opinion mining and sentiment analysis*, *Found. Trends® Inf. Retr.* 2 (2008) 1–135.
6. A. Kumar, T. Sebastian, *Sentiment analysis: A perspective on its past, present and future*, *Int. J. Intell. Syst. Appl.* 4 (2012) 1–14.
7. D. Li, Z. Luo, Y. Ding, J. Tang, G.G.-Z. Sun, X. Dai, J. Du, J. Zhang, S. Kong, *User-level microblogging recommendation incorporating social influence*, *J. Assoc. Inf. Sci. Technol.* 68 (2017) 553–568.
8. A. Pak, P. Paroubek, *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, *Int. Lang. Resour. Eval.* (2010) 1320–1326.
9. B. Liu, *Sentiment Analysis Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, 2015.
10. A. Kumar, A. Joshi, *Ontology Driven Sentiment Analysis on Social Web for Government Intelligence*, *Spec. Collect. eGovernment Innov. India.* (2017) 134–139.
11. A. Kumar, R. Khorwal, S. Chaudhary, *A Survey on Sentiment Analysis using Swarm Intelligence*, *Indian J. Sci. Technol.* 9 (2016) 1–7.
12. N.K. Sinha, M.M. Gupta, L.A. Zadeh, *Soft Computing and Intelligent Systems, Theory and Applications*, 2000.
13. S.N. Sivanandam, S.N. Deepa, *Principles of Soft Computing*, Wiley India first ed., 2007.
14. V.E. Balas, J. Fodor, A.R. Várkonyi-Kóczy, *New Concepts and Applications in Soft Computing*, 2013.
15. M. Tsytsarau, T. Palpanas, *Survey on mining subjective data on the web*, *Data Min. Knowl. Discov.* 24 (2012) 478–514.
16. E. Cambria, B. Schuller, Y. Xia, C. Havasi, *New Avenues in Opinion Mining and Sentiment Analysis*, *IEEE Intell. Syst.* 28 (2013) 15–21.
17. R. Feldman, *Techniques and applications for sentiment analysis*, *Commun. ACM.* 56 (2013) 82–89.
18. A. Montoyo, P. Martínez-Barco, A. Balahur, *An overview of the current state of the area and envisaged developments*, *Decis. Support Syst.* 53 (2012) 675–679.

19. W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, *Ain Shams Eng. J.* 5 (2014) 1093–1113.
20. K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: Opinion extraction and semantic classification of product reviews, *12th Int. Conf. World Wide Web.* (2003) 519–528.
21. W. Guan, H. Gao, M. Yang, Y. Li, H. Ma, W. Qian, Z. Cao, X. Yang, Analyzing user behavior of the micro-blogging website Sina Weibo during hot social events, *Phys. A Stat. Mech. Its Appl.* 395 (2014) 340–351.
22. A. García, Global financial indices and Twitter sentiment: A random matrix theory approach, *Phys. A Stat. Mech. Its Appl.* 461 (2016) 509–522.
23. Bhatia, M. P. S., & Kumar, A. (2008). Information retrieval and machine learning: supporting technologies for web mining research and practice. *Webology*, 5(2), 5.
24. S. Finn, E. Mustafaraj, Learning to discover political activism in the Twitterverse, *KI-Künstliche Intelligenz.* 27 (2013) 17–24.
25. Kumar, A., & Khorwal, R. (2017). Firefly algorithm for feature selection in sentiment analysis. In *Computational Intelligence in Data Mining* (pp. 693-703). Springer, Singapore.
26. B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, *Tech. Rep. EBSE.* 1 (2007) 1–57.
27. S.R. Yerva, Z. Miklós, K. Aberer, Quality-aware similarity assessment for entity matching in Web data, *Inf. Syst.* 37 (2012) 336–351.
28. T. Lou, J. Tang, J. Hopcroft, Z. Fang, X. Ding, Learning to predict reciprocity and triadic closure in social networks, *ACM Trans. Knowl. Discov. from Data.* 7 (2013) 1–25.
29. M. Arias, A. Arratia, R. Xuriguera, Forecasting with Twitter data, *ACM Trans. Intell. Syst. Technol.* 5 (2013).
30. A. Trilla, F. Alias, Sentence-based sentiment analysis for expressive text-to-speech, *IEEE Trans. Audio. Speech. Lang. Processing.* 21 (2013) 223–233.
31. S. Tuarob, C.S. Tucker, M. Salathe, N. Ram, An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages, *J. Biomed. Inform.* 49 (2014) 255–268.
32. M. Morchid, R. Dufour, P.-M. Bousquet, G. Linarès, J.-M. Torres-Moreno, Feature selection using Principal Component Analysis for massive retweet detection, *Pattern Recognit. Lett.* 49 (2014) 33–39.
33. A. Montejo-Ráez, E. Martínez-Cámara, M.T. Martín-Valdivia, L.A. Ureña-López, Ranked wordnet graph for sentiment polarity classification in Twitter, *Comput. Speech Lang.* 28 (2014).
34. J. Smailović, M. Grčar, N. Lavrač, M. Žnidaršič, Stream-based active learning for sentiment analysis in the financial domain, *Inf. Sci. (Ny).* 285 (2014) 181–203.
35. G. Boella, L. Di Caro, A. Ruggeri, L. Robaldo, Learning from syntax generalizations for automatic semantic annotation, *J. Intell. Inf. Syst.* 43 (2014) 231–246.

36. J. Brynielsson, F. Johansson, C. Jonsson, A. Westling, Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises, *Secur. Inform.* 3 (2014) 1–11.
37. Y. Arakawa, A. Kameda, A. Aizawa, T. Suzuki, Adding Twitter-specific features to stylistic features for classifying tweets by user type and number of retweets, *J. Assoc. Inf. Sci. Technol.* 65 (2014) 1416–1423.
38. P. Burnap, M.L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, A. Voss, Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack, *Soc. Netw. Anal. Min.* 4 (2014) 1–14.
39. A. Makazhanov, D. Rafiei, M. Waqar, Predicting political preference of Twitter users, *Soc. Netw. Anal. Min.* 4 (2014) 1–15.
40. P. Bogdanov, M. Busch, J. Moehlis, A.K. Singh, B.K. Szymanski, Modeling individual topic-specific behavior and influence backbone networks in social media, *Soc. Netw. Anal. Min.* 4 (2014) 1–16.
41. Y.-R. Lin, D. Margolin, The ripple of fear, sympathy and solidarity during the Boston bombings, *EPJ Data Sci.* 3 (2014).
42. X. Fu, Y. Shen, Study of collective user behaviour in Twitter: a fuzzy approach, *Neural Comput. Appl.* 25 (2014) 1603–1614.
43. X. Chen, M. Vorvoreanu, K. Madhavan, Mining social media data for understanding students' learning experiences, *IEEE Trans. Learn. Technol.* 7 (2014) 246–259.
44. S. Liu, X. Cheng, F. Li, F. Li, TASC: topic-adaptive sentiment classification on dynamic tweets, *IEEE Trans. Knowl. Data Eng.* 27 (2015) 1696–1709.
45. J. Kranjc, J. Smailović, V. Podpečan, M. Grčar, M. Žnidaršič, N. Lavrač, Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform, *Inf. Process. Manag.* 51 (2015) 187–203.
46. B. Sluban, J. Smailović, S. Battiston, I. Mozetič, Sentiment leaning of influential communities in social networks, *Comput. Soc. Networks.* 2 (2015) 1–21.
47. P. Burnap, M.L. Williams, Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making, *Policy & Internet.* 7 (2015) 223–242.
48. A. Zubiaga, D. Spina, R. Martinez, V. Fresno, Real-time classification of Twitter trends, *J. Assoc. Inf. Sci. Technol.* 66 (2015) 462–473.
49. W. Magdy, H. Sajjad, T. El-Ganainy, F. Sebastiani, Bridging social media via distant supervision, *Soc. Netw. Anal. Min.* 5 (2015) 1–12.
50. M. Tsytsarau, T. Palpanas, Managing Diverse Sentiments at Large Scale, *IEEE Trans. Knowl. Data Eng.* 28 (2016) 3028–3040.
51. P. Andriotis, G. Oikonomou, T. Tryfonas, S. Li, Highlighting relationships of a smartphone's social ecosystem in potentially large investigations, *IEEE Trans. Cybern.* 46 (2016) 1974–1985.
52. D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, M. Zhou, Sentiment embeddings with applications to sentiment analysis, *IEEE Trans. Knowl. Data Eng.* 28 (2016) 496–509.

53. M.-H. Peetz, M. de Rijke, R. Kaptein, Estimating reputation polarity on microblog posts, *Inf. Process. Manag.* 52 (2016) 193–216.
54. E. Sulis, D.I.H. Farías, P. Rosso, V. Patti, G. Ruffo, Figurative messages and affect in Twitter: Differences between# irony,# sarcasm and# not, *Knowledge-Based Syst.* 108 (2016) 132–143.
55. F. Wu, Y. Song, Y. Huang, Microblog sentiment classification with heterogeneous sentiment knowledge, *Inf. Sci. (Ny)*. 373 (2016) 149–164.
56. S.L. Lo, R. Chiong, D. Cornforth, Ranking of high-value social audiences on Twitter, *Decis. Support Syst.* 85 (2016) 34–48.
57. W. van Zoonen, G.L.A. Toni, Social media research: The application of supervised machine learning in organizational communication research, *Comput. Human Behav.* 63 (2016) 132–141.
58. Z. Wang, X. Cui, L. Gao, Q. Yin, L. Ke, S. Zhang, "A hybrid model of sentimental entity recognition on mobile social media, *EURASIP J. Wirel. Commun. Netw.* 1 (2016) 253.
59. F. Celli, A. Ghosh, F. Alam, G. Riccardi, In the mood for sharing contents: Emotions, personality and interaction styles in the diffusion of news, *Inf. Process. Manag.* 52 (2016) 93–98.
60. R.A. Igawa, S. Barbon Jr, K.C.S. Paulo, G.S. Kido, R.C. Guido, M.L. Proença Júnior, I.N. da Silva, Account classification in online social networks with lbca and wavelets, *Inf. Sci. (Ny)*. 332 (2016) 72–83.
61. G. Korkmaz, J. Cadena, C.J. Kuhlman, A. Marathe, A. Vullikanti, N. Ramakrishnan, Multi-source models for civil unrest forecasting, *Soc. Netw. Anal. Min.* 6 (2016) 1–25.
62. P. Burnap, M.L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics, *EPJ Data Sci.* 5 (2016) 11.
63. N. Oliveira, P. Cortez, N. Areal, The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices, *Expert Syst. Appl.* 73 (2017) 125–144.
64. I. Perikos, I. Hatzilygeroudis, Recognizing emotions in text using ensemble of classifiers, *Eng. Appl. Artif. Intell.* 51 (2016) 191–201.
65. M.L. Brocardo, I. Traore, I. Woungang, M.S. Obaidat, Authorship verification using deep belief network systems, *Int. J. Commun. Syst.* (2017) 1–10.
66. M. Bouazizi, T.O. Ohtsuki, A Pattern-Based Approach for Sarcasm Detection on Twitter, *IEEE Access.* 4 (2016) 5477–5488.
67. D.I.H. Farías, V. Patti, P. Rosso, Irony detection in Twitter: The role of affective content, *ACM Trans. Internet Technol.* 16 (2016) 19.
68. V. Sintsova, P. Pu, Dystemo: Distant Supervision Method for Multi-Category Emotion Recognition in Tweets, *ACM Trans. Intell. Syst. Technol.* 8 (2016) 1–22.
69. L.R. Nair, S.D. Shetty, S.D. Shetty, Applying spark based machine learning model on streaming big data for health status prediction, *Comput. Electr. Eng.* (2017) 1–7.
70. L. Cui, X. Zhang, A.K. Qin, T. Sellis, L. Wu, CDS: Collaborative distant supervision for Twitter account classification, *Expert Syst. Appl.* 83 (2017) 94–103.

71. P. Pérez-Gállego, J.R. Quevedo, J.J. del Coz, Using ensembles for problems with characterizable changes in data distribution: A case study on quantification, *Inf. Fusion*. 34 (2017) 87–100.
72. T. Alsinet, J. Argelich, R. Béjar, C. Fernández, C. Mateu, J. Planes, Weighted argumentation for analysis of discussions in Twitter, *Int. J. Approx. Reason.* 85 (2017) 21–35.
73. Z. Jianqiang, G. Xiaolin, Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis, *IEEE Access*. 5 (2017) 2870–2879.
74. Jain VK, Kumar S. Effective surveillance and predictive mapping of mosquito-borne diseases using social media. *Journal of Computational Science*. 2018; 25: 406-415.
75. Keshavarz H, Abadeh MS. ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems*. 2017; 122: 1-16.
76. Xiong S, Lv H, Zhao W, Ji D. Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*. 2018; 275: 2459-2466.
77. Neppalli VK, Caragea C, Squicciarini A, Tapia A, Stehle S. Sentiment analysis during Hurricane Sandy in emergency response. *International journal of disaster risk reduction*. 2018; 21: 213-222.
78. Singh A, Shukla N, Mishra N. Social media data analytics to improve supply chain management in food industries. *Transportation Research Part E: Logistics and Transportation Review*. 2018; 114: 398-415.
79. Xiaomei Z, Jing Y, Jianpei Z, Hongyu H. Microblog sentiment analysis with weak dependency connections. *Knowledge-Based Systems*. 2018; 142: 170-180.
80. Khan I, Naqvi SK, Alam M, Rizvi SNA. An efficient framework for real-time tweet classification. *International Journal of Information Technology*. 2017; 9(2): 215-221.
81. Bouazizi M, Ohtsuki T. A pattern-based approach for multi-class sentiment analysis in Twitter. *IEEE Access*. 2017; 5: 20617-20639.
82. Li B, Chan KC, Ou C, Ruifeng S. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. *Information Systems*. 2017; 69: 81-92.
83. Jianqiang Z, Xiaolin G, Xuejun Z. Deep Convolution Neural Networks for Twitter Sentiment Analysis. *IEEE Access*. 2018; 6: 23253-23260.
84. Ghiassi M, Lee S. A domain transferable lexicon set for Twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*. 2018; 106: 197-216.
85. Symeonidis S, Effrosynidis D, Arampatzis A. A Comparative Evaluation of Pre-Processing Techniques and their Interactions for Twitter Sentiment Analysis. *Expert Systems with Applications*. 2018; 110: 298-310.
86. Brendano, GitHub.com. <https://github.com/brendano/ark-tweet-nlp/tree/master/src/cmu/arktweetnlp> (accessed 2 January 2018).
87. Shahana, P.H., Omman, B., 2015. Evaluation of Features on Sentimental Analysis. *Procedia Computer Science, Elsevier*. 46, 1585-1592.

88. Rosenthal, S., Farra, N. and Nakov, P. SemEval-2017 task 4: Sentiment analysis in Twitter. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association of Computational Linguistics, 502-518 (2017).
89. SemEval-2017 Task 4, Sentiment Analysis in Twitter. <http://alt.qcri.org/semeval2017/task4/> (accessed 2 January 2018).
90. G. Wang, et al. Sentiment classification: The contribution of ensemble learning. *Decision Support Systems* (2013), <http://dx.doi.org/10.1016/j.dss.2013.08.002>
91. Kanakaraj, M., and Guddeti, R. M. R. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In: IEEE International Conference on Semantic Computing, pp. 169-170 (2015).
92. Catal, C., and Nangir, M. A sentiment classification model based on multiple classifiers. *Applied Soft Computing* 50, 135-141(2017).
93. Wan, Y., and Gao, Q. An ensemble sentiment classification system of Twitter data for airline services analysis. In: IEEE International Conference on Data Mining, pp. 1318-1325 (2015).
94. Athanasiou, V.; Maragoudakis, M. A Novel, Gradient Boosting Framework for Sentiment Analysis in Languages where NLP Resources Are Not Plentiful: A Case Study for Modern Greek. *Algorithms* 10(34) (2017).
95. Wijesinghe, Isuru. *Sentiment Analysis* (2015)
96. Geurts, Pierre, Ernst, Damien and Wehenkel, Louis. Extremely Randomized Trees. *Machine Learning* 63, 3-42 (2006). 10.1007/s10994-006-6226-1.
97. Mirjalili, S., Mirjalili, S.M., 2014. A. Lewis, Grey wolf optimizer. *Advances in Engineering Software*. 69, 46-61.
98. Mirjalili, S., 2015. Moth-Flame Optimization Algorithm: A Novel Nature-inspired Heuristic Paradigm. *Knowledge-Based Systems*. 89, 228-249.
99. Omar, N., Jusoh, F., Ibrahim R., et. al. Review of Feature Selection for Solving Classification Problems. *Journal of Information System Research and Innovation* 3, 64-70 (2013).
100. Omar, N., Othman, M.S. Particle Swarm Optimization Feature Selection for Classification of Survival Analysis in Cancer. *International Journal of Innovative Computing* 2(1) (2013).
101. Shi, Y., Eberhart, R.C. A modified particle swarm optimizer. In: Proc. IEEE Int. Conf. Evolutionary Computation, Anchorage, AK, USA, pp. 69-73 (1998).
102. J. Pennington, R. Socher and C. Manning, "Glove: Global vectors for word representation", In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543, 2014.
103. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities". *Proceedings of the National Academy of Sciences* 79 (8): 2554, 1982.
104. T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing". *IEEE Computational Intelligence magazine*, vol. 13, no. 3, pp. 55-75, 2017.

- 105.Z. Reznikova, "Animal Intelligence from Individual to Social Cognition". Cambridge University Press, 2007.
- 106.B. Ryabko & Z. Reznikova, "The use of ideas of information theory for studying 'language' and intelligence in ants". *Entropy* 11 (4): 836–853, 2009.
- 107.H. Hui-Huang C.W. Hsieh, & M.D. Lu, "Hybrid feature selection by combining filters and wrappers". *Expert Syst. Appl*, vol. 38, pp. 8144-8150, 2011.
- 108.R. Tang, S. Fong, X. S.Yang & S. Deb, "Wolf search algorithm with ephemeral memory". In *Seventh International Conference on Digital Information Management (ICDIM 2012)*, IEEE, 2012, pp. 165-172.

# Appendix-A

---

---

## *List of Publications*



## LIST OF PUBLICATIONS

### Journal(s)

1. Kumar, A. and Jaiswal, A. A Deep Swarm Optimized Model for leveraging Industrial Data Analytics in Cognitive Manufacturing. *IEEE Transactions on Industrial Informatics*. 2020 Jun 29. **[SCIE JOURNAL, Impact factor: 9.112]**. DOI: 10.1109/TII.2020.3005532.
2. Kumar, A. and Jaiswal, A. Systematic literature review of sentiment analysis on Twitter using soft computing techniques. *Concurrency and Computation: Practice and Experience, Wiley*. 2020 Jan 10:32(1):e5107. **[SCIE JOURNAL, Impact factor: 1.447]**. <https://doi.org/10.1002/cpe.5107>.
3. Kumar, A. and Jaiswal, A. Swarm Intelligence Based Optimal Feature Selection for Enhanced Predictive Sentiment Accuracy on Twitter. *Multimedia Tools and Applications, Springer*. 2019 Oct 1; 78(20):29529-53. **[SCIE JOURNAL, Impact factor: 2.313]**. <https://doi.org/10.1007/s11042-019-7278-0>.
4. Kumar, A. and Jaiswal, A. Scalable Intelligent Data-Driven Decision Making for Cognitive Cities. *Energy Systems (S.I.: Energy efficiency in building using intelligent computing for smart cities), Springer*. 2019 Nov 19: 1-9. **[SCImago, SCOPUS JOURNAL, Impact factor: 1.65]**. <https://doi.org/10.1007/s12667-019-00369-5>.
5. Kumar, A. and Jaiswal, A. Deep Network Learning Based Sentiment Classification on User-generated Big Data. Recent Patents on Computer Science, *Bentham Science*, 2019.12:1. **[SCOPUS] JOURNAL**. <https://doi.org/10.2174/2213275912666190409152308>.

### Conference(s)

1. Kumar, A. and Jaiswal, A. Empirical Study of Twitter and Tumblr for sentiment analysis using soft computing techniques. *In Proceedings of the World Congress on Engineering and Computer Science*, 2017. Vol. 1, pp. 1-5. **[SCOPUS]**. 978-988-14047-5-6.
2. Kumar, A. and Jaiswal, A. Particle Swarm Optimization-based Ensemble Learning for Enhanced Predictive Sentiment Accuracy of Online-micro Tweets. *In Proceedings of International Conference on Emerging Trends in Information Technology, Springer, 2019*. Pp. 633-646. **[SCOPUS]**. DOI: 10.1007/978-3-030-30577-2\_5