BIO-INSPIRED COMPUTING FOR OUTLIER DETECTION: SELECT STUDIES IN WEB 3.0 DOMAIN

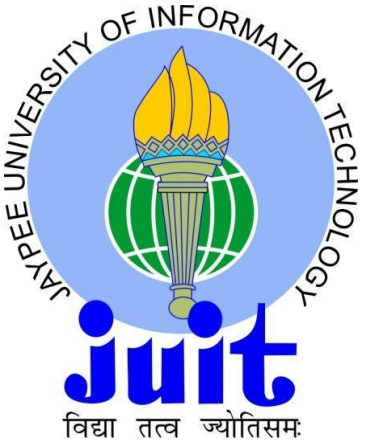
Thesis submitted in fulfilment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

BY

REEMA ASWANI

136213



Under the supervision of

PROF. S. P. GHRERA

DR. SATISH CHANDRA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT, SOLAN-173234, HIMACHAL PRADESH, INDIA

JANUARY, 2019

TABLE OF CONTENTS

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled **"Bio-Inspired Computing for Outlier Detection: Select Studies in Web 3.0 Domain"** submitted at **Jaypee University of Information Technology, Waknaghat, India**, is an authentic record of my work carried out under the supervision of **Prof. Dr. S. P. Ghrera** and **Dr. Satish Chandra**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Thesis.

(Signature of the Scholar)
(Reema Aswani)
Department of Computer Science and Engineering
Jaypee University of Information Technology, Waknaghat, India
Date ( )

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled **"Bio-Inspired Computing for Outlier Detection: Select Studies in Web 3.0 Domain"**, submitted by **Reema Aswani** at **Jaypee University of Information Technology, Waknaghat**, **India,** is a bonafide record of her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Signature of Supervisor)
(Prof. S. P. Ghrera)
Department of Computer Science and Engineering
Jaypee University of Information and Technology
Waknaghat, India (173234)
Date ( )

(Signature of Supervisor)
(Dr. Satish Chandra)
Department of Computer Science and Engineering
Jaypee Institute of Information and Technology
Noida, India (201309)
Date ( )

ACKNOWLEDGEMENT

I express my profound and sincere gratitude and appreciation to my supervisor Prof. S. P. Ghrera and my co-supervisor Dr. Satish Chandra for their continuous and holistic valuable guidance in every step. Their guidance has helped me realize the research problems, the salient issues and towards shaping this PhD thesis from disjoined thoughts. I am fully indebted for their personal and honest support and for their calibrated professional guidance in many crucial areas of research works including analysis of data, its transparent interpretation and techniques of presentation of the facts in a structured and methodical format. I have been fortunate to have advisors who gave me the freedom to explore on my own, and at the same time provided the guidance to recover when my steps faltered. Their patience and support helped me overcome many crisis situations and finish this dissertation.

I further take the opportunity to express my sincere and honest gratitude to Dr. Arpan Kumar Kar from Indian Institute of Technology, Delhi for his valuable, scholarly and meaningful inputs which did facilitate and galvanize my research work in a very methodical and calibrated manner. By his efficient and sincere efforts, he also provided me a precious environment conducive to help in evolving and familiarizing the emerging domain of bio-inspired computing and social media analytics helpful for this research study.

I express my gratitude to the DPMC members, Prof. S. K. Khah, Dr. Amit Kr. Singh, Dr. Yugal Kumar and Dr. Rajni Mohana for effective and critical inputs wherever necessary for the improvisation of the work. I am grateful to all of them for the long discussions that helped me sort out the technical details of my work. I am also indebted to my professional colleagues for their constant support and for providing me opportunities to conduct this research works. I also express my honest gratitude to my friends Aastha Modgil, Oshin Sharma and Ruhi Mahajan for my pleasant stay at JUIT during the initial days of my work.

Finally, things will remain incomplete if I do not express my gratitude and obligations to Dr. Vivek Sehgal for being a reliable point of contact and for the encouraging the use of proper timelines. Further, I appreciate academic support of Jaypee University of Information Technology, Waknaghat specially Mr. Amit Kumar Srivastava for his continuous cooperation and sincere assistance.

I would also like to express gratitude to my closest friend Nishtha Ahuja who has been a constant support throughout the journey from my visits to JUIT to long motivation phone calls discussing every little detail without showing any signs of boredom.

Most importantly, none of this would have been possible without the love and patience of my family, my father Mr. Mukesh Aswani, my mother Mrs. Rita Aswani, my grandmother Mrs. Sarla Aswani and my younger brother Hitesh Aswani. They have been a constant source of love, concern, support and strength all these years.

Last but not the least, this dissertation would not have been possible without the most special person in my life, my husband Arjun Assi, who has always been with me through thick and thin. I am grateful to him for believing in me and for all the efforts he has made to make this happen. He has been a constant pillar of strength and has helped me in every way possible.

(Reema Aswani)
Department of Computer Science and Engineering
Jaypee University of Information Technology, Waknaghat, India

DEDICATION

I dedicate this dissertation to all my loved ones, those with me today and those who have passed on.

ABSTRACT

In the current scenario, data analytics has emerged as an inevitable domain. Increasing magnitude of data not only in terms of volume but also variety and veracity has made the subsequent analysis and decision making a challenging task. Researches and practitioners have adopted variety of data analytics approaches and frameworks for retrieving useful information from data of such magnitude. Several data mining and information retrieval tools have been designed to address this huge data inflow and associated problems.

The entire business intelligence can actually go futile if the available data is not in the correct format or comprises of aberrations/outliers. Data outliers are nothing but instances lying away from majority of available instances. Thus, these data points become of particular interest to researchers working in the field of data analytics. These data instances may occur due to errors made while acquiring the data, data variations or some deviations in the data itself that result into abnormalities. This makes outlier detection an inevitable step for efficient and effective information retrieval.

Further, advances in the domain of information technology have increased exponentially with the rising growth in the use of the internet gradually generating innovation in diverse domains. This leads to the emergence of Web 3.0 with huge amount of data being generated from social media and other interactive web platforms. Thus, the contribution of this work is twofold, both methodological as well as application oriented focusing on the domain of Web 3.0. The work targets select studies in the Web 3.0 domain highlighting outliers of different types. Methodologically, the work proposes several hybrid bio inspired computing algorithms by integrating them with traditional k-Means and k-nearest neighbor algorithms.

The bio-inspired computing algorithms are known to produce promising results when compared to traditional machine learning algorithms that are usually utilized for outlier detection. This work thus leverages the methodological advantage of these approaches and applies them to target a less frequent application of these algorithms. The select studies thus use the proposed

hybrid bio inspired approaches for outlier detection in relevant studies of Web 3.0. The work is focused on three research problems in the Web 3.0 domain including search engine marketing, social media marketing and influencer marketing. The use of hybrid bio-inspired computing algorithms eliminates locally optimum solutions and catalyzes the convergence of the solution. The outlier definitions vary with the changing Web 3.0 problem under consideration. The findings of the work have policy implications in domains of e-commerce, social media and influencer marketing.

LIST OF ACRONYMS AND ABBREVIATIONS

ABC	Artificial Bee Colony
AdaBoost	Adaptive Boosting
AR	Alexa Rank
BA	Bat Algorithm
CCS	Chaotic Cuckoo Search
CF	Citation Flow
CLARA	Clustering LARge Applications
CLARANS	Clustering Large Applications based on RANdomized Search
COA	Chaotic Optimization Algorithms
CSA	Cuckoo Search Algorithm
DA	Domain Authority
EBL	External Back Links
ELL	External Equity Links
FA	Firefly Algorithm
GWO	Grey Wolf Optimizer
HITS	Hyperlink-Induced Topic Search
HL	SemRush Hostname Links
KNN	K-Nearest Neighbor
LDA	Linear Discriminant Analysis
LI	Links In
MR	Moz Rank
NB	Naïve Bayes
NFL	No Free Lunch
NP	Non-deterministic Polynomial-time
PA	Page Authority
PAM	Partitioning Around Medoids
RD	Referred Domains

RF	Random Forest
Rpart	Recursive Partitioning and Regression Trees
SD	Standard Deviation
SEM	Search Engine Marketing
SEO	Search Engine Optimization
SERP	Search Engine Results Page
SMO	Sequential Minimal Optimization
SVM	Support Vector Machine
TF	Trust Flow
UGC	User Generated Content
UL	SemRush URL Links
WSA	Wolf Search Algorithm

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

1. INTRODUCTION

1.1 Problem Statement

Advances in the domain of information technology, specifically with a focus on communication and multimedia, have increased exponentially with the rising growth in the use of the internet. These advances are gradually opening innovative avenues in diverse domains leading to the expansion of new business practices based on the knowledge and availability of information [1]. This is where the importance of networks and partnerships between organizations and their agents plays a crucial role. Advancement in Web 3.0 technologies has completely revamped organization structures and value networks with a possibility of decision-making processes. The efficient use of the interactive web is thus crucial in current environment in order to consolidate the advantages of these businesses practices [2].

Further, the Web 3.0 primarily comprises of popular applications including Facebook, MySpace, Twitter and YouTube. These platforms have found interesting mechanisms to disseminate content and deliver the functionality commonly referred to as the "read-write Web". These platforms are also known to cater to masses having a greater overall impact. The domain of Web 3.0 has thus gained immense popularity with increased use of internet by individuals and organizations. The usage in the current scenario is not restricted to mere communication but targets various aspects of digital marketing. This engagement and interaction on platforms generates large amount of User Generated Content (both structured and unstructured). However, the explosion of data comes with problems of its own, and path breaking applications for this new generation of technology are yet to be developed and adopted by the users [3].

This work caters both to the domain and methodological contribution. On one hand we understand the problems faced in the Web 3.0 domain. Outlier detection on the other hand, is a widely explored area in various domains including wireless sensor networks [4], medical data

[5] and fraud detection [6] to name a few. Literature highlights various outlier detection techniques and extensive reviews on methodologies [7] [8]. These approaches are categorized into nearest neighbor, statistical, clustering and classification based to name a few [9]. K-Means and KNN are one of the most popular algorithms that can be used for clustering datasets that is often used for outlier detection [10] [11]. The only pitfall of this approach is that it often falls into local optima. Further, due to increased variety and volume of the data, it has become computationally complex to analyze it using simple machine learning approaches. Thus, due to these issues in heuristic approaches, meta-heuristic approaches have become prominent over the decades [12] [13].

1.2 Motivation and Contribution

The traditional machine learning algorithms have been extensively used for outlier detection. However, these algorithms are often prone to fall in local optima. K-Means and k-nearest neighbors are often the most common choices when it comes to clustering and classification datasets that is often used for outlier. However, these traditional approaches are prone to fall in local optima. Further, due to increased variety and volume of the data, it has become computationally complex to analyze it using simple machine learning approaches. Therefore, keeping the said concerns surrounding heuristic approaches in mind, meta-heuristic approaches have become prominent over the decades.

Bio-inspired computing algorithms are known overcome local optima and are popular for converging to a globally optimum solution. Existing studies in literature highlight that there are hardly any studies in outlier detection using bio inspired computing approaches. It has been noted that for the past many decades bio inspired computing and optimization techniques have been successful in solving these dynamic and highly complex problems. This has given rise to a variety of nature inspired techniques for computing as discussed by Kar [14] in an extensive review. The inflow of these techniques started way back in the 1970s but not all of these became prominent. The techniques comprised of neural networks [15], genetic algorithms [16], and leaping frog algorithm [17] [18]. Then the famous ant colony optimization [19] and particle swarm optimization [20] which put a full stop on the heuristic based approaches. Meta heuristics

became the new trend in bio inspired computing, within which algorithms like bacterial foraging optimization [21], Cuckoo Search [22], artificial bee colony [23], firefly algorithm [24], bat algorithm [25] and grey wolf optimization [26] became popular. In this work, we explore the application of these meta-heuristic approaches for detecting outliers.

Further, with the era of digitization and the emergence of Web 3.0 there is increased use of social media and search engines. This results in the generation of large amount of data, both structured and non-structured including UGC. With this sudden explosion of data generation which has increased the probability of erroneous/anomalous data. This has generated a requirement of getting optimal solutions for data driven problems. It has become a great challenge to find a solution best suited for the given data driven issues. The data contains critical information which can have implications in the domains of e-commerce, public policy, search engine marketing, and influencer marketing to name a few. This data if mined and analyzed properly can give useful insights to be used in the above mentioned domains.

Thus, the contribution of this work is twofold, both methodological as well as application oriented focusing on the domain of Web 3.0. The work targets select studies in the Web 3.0 domain highlighting outliers of different types. Methodologically, the work proposes several hybrid bio inspired computing algorithms by integrating them with traditional k-Means and k-nearest neighbors algorithms. The select studies thus use the proposed hybrid bio inspired approaches for outlier detection in relevant studies of Web 3.0. The work is focused on the three research problems in the Web 3.0 domain including search engine marketing, social media marketing and influencer marketing. The subsequent use of hybrid bio-inspired computing algorithms eliminates locally optimum solutions and catalyzes the convergence of the solution. The findings of the study have policy implications in domains of e-commerce, social media and influencer marketing.

1.3 Research Problems

Bio-inspired computing algorithms are known for optimizing an objective function either for minimization or maximization. These meta-heuristic approaches when integrated with traditional machine learning algorithms help in avoiding locally optimal solutions expedite the convergence to the solution and enhance the accuracy of the obtained solution. In addition to the methodological contribution, the work addresses concerns in the domain of Web 3.0 surrounding outliers. To address the concerns surrounding detection/identification of outliers in the Web 3.0 domain the study proposes three research objectives with sub-problems having practical implications in the industry.

The research objectives attempt to address practical problems faced by organizations and individuals with the advent of Web 3.0 and increased use of internet in every facet of communication and engagement. The findings of this work have policy implications in domains of e-commerce, social media and influencer marketing. The objectives include outlier detection in supervised domain, unsupervised domain and integration chaos theory for outlier detection. A brief description of these is discussed subsequently.

1.3.1 Outlier Detection in Supervised Scenario

The first problem explores the use of hybrid bio-inspired computing algorithms for detecting outliers in the supervised scenario. The supervised scenario may comprise of classification or regression datasets wherein we have a continuous or nominal output label. The motivation behind these studies is to avoid locally optimum solutions and minimize convergence iterations which are significantly high when it comes to traditional machine learning approaches implemented for big datasets. The sections implements hybrid grey wolf optimizer, wolf search algorithm and bat algorithm. The latter have been explored for detecting outliers among influencer blogs.

1.3.2 Outlier Detection in Unsupervised Scenario

This section of thesis is surrounding case scenarios surrounding outlier detection in an unsupervised scenario where the dataset does not have an output label for outliers. Clustering approaches are often common to deal to mine relevant information from such datasets. These approaches cluster similar data points together to reflecting appropriate patterns. The algorithms proposed for this chapter include hybrid artificial bee colony and grey wolf optimization along with k-nearest neighbors. The section considers social media datasets from Twitter for identifying outliers in the context of buzz and fake profiles.

1.3.3 Integrating Chaos in Outlier Detection

This sections attempts to improvise on the convergence speed to reach to the solution while detecting outliers by exploring a better search space while optimizing through the uses of chaos theory. Almost every meta-heuristic algorithm with stochastic components achieves randomness through probability distributions. The chaos theory uses variables in random-based optimization and is known to conduct the overall search at higher speeds, the main reason for the same is often non-repetition of chaos [27]. Due to these properties of chaos theory, algorithms can conduct iterative search at faster than standard search possessing standard distributions.

1.4 Performance Metrics

This section provides a brief description of the metrics that have been used to validate, compare and contrast the proposed approaches in the thesis.

Accuracy: It is a measure of the closeness to the true value. Mathematically computed using true positive (*tp*), false positive (*fp*), true negative (*tn*) and false negative (*fn*).

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Threshold: Threshold is a numeric value for every attribute, beyond which the data point is considered as an outlier. The value may vary with the dataset under consideration depending on the domain and data distribution.

Classification error rate: In statistical classification, classification error rate of a random outcome and is analogous to the irreducible error.

Precision: It is also referred to as a positive predictive value and is the ratio of positive values [true positive (*tp*)] to the total retrieved values [sum of true positive and false positive (*fp*)].

$$\text{Precision} = \frac{tp}{tp + fp}$$

Recall: It is often referred to as sensitivity and is the ratio of the positive values [true positives (*tp*)] to the total number of relevant values retrieved [sum of true positive (*tp*) and false negative (*fp*)]

$$\text{Recall} = \frac{tp}{tp + fn}$$

Running Time: It is the time required for the algorithm to run and produce results. The current work represents the running time in seconds.

P-value: It is the calculated probability of finding the observed results. In statistical analyses P-values are often used as a test for significance.

Significance Level: Usually, an alpha value of 0.05 is used as the cutoff for determining significance. If the p-value is less than 0.05, we can conclude that a significant difference exists.

Spam Score: A measure to gauge whether an instance is a spam/outlier instance or not. Threshold of being classified as spam varies base d on the problem under consideration.

Objective Function: Usually used in linear programming problem to decide the objective of the problem which can be either a minimization or maximization. The outlier detection problem is modeled as an objective function in this work.

Cluster Centers: The central value of the attributes that are being modeled in the work into clusters of outlier and non-outlier instances depending on the problem domain.

F-Measure: is a measure of a accuracy of a test and uses combination of precision and recall. It does not take into consideration the true negatives.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Iteration Scores: The scores for finalizing the attributes to be considered for final analysis using a Delphi approach which uses a voting mechanism to reach to a consensus.

Convergence Iterations: This is a count of the number of iterations which the algorithm requires to reach to an optimal solution. Usually referred to as the iteration at which the value of objective function stabilizes.

1.5 Thesis Outline

The thesis is systematically organized into six chapters based on the research problems identified. Chapter 1 is an introduction to the work comprising of the problem statement, the motivation for conducting the research and the subsequent contribution. It further briefly defines the research problems and provides a brief description of the performance metrics used to validate and compare various proposed approaches. Chapter 2 is primarily focused on data description. This chapter briefly describes the metrics for each dataset that is modeled for identification of outliers in the form of influencer blogs, content buzz and fake profiles amongst others. The Chapter 3 of the thesis is a detailed literature review on the three primary domains on which the work is focused. The first being traditional outlier detection algorithms followed by a section on existing outlier detection approaches using meta-heuristic approaches specifically bio-inspired computing followed by an introduction into the emerging domain of Web 3.0 and existing studies surrounding the same. The subsequent three chapters are completely focused on the research problems as identified in the introduction. Chapter 4 addresses outlier detection for supervised scenario comprising of two sub-problems. The first sub-section proposes a hybrid grey wolf optimization approach for mining outliers for regression and classification datasets. The second sub-problem focuses on the identification of influencer blogs as outliers using proposed wolf search algorithm and bat algorithm. Chapter 5 attempts to focus on outlier detection in the unsupervised scenario where the datasets do not possess an output label. The sub-problems addressed in this chapter pertain specifically on Twitter. The outliers mined are in the form of content buzz and fake profiles on the popular micro-blogging platform. The approaches used comprise of integrated artificial bee colony and

grey wolf optimizer. The Chapter 6 provides interesting insights by integrated chaos theory with bio-inspired approaches. Chaotic variants of k-Means integrated cuckoo search and firefly algorithm are proposed. The sub-problems mine outliers in the form of popular online content and spam websites for search engine marketing. Lastly, Chapter 7 provides the concluding discussions and the future scope of the work done.

CHAPTER 2

2. LITERATURE REVIEW

This chapter highlights the literature surrounding the three key themes of the current work. Evidences from existing studies surrounding outlier detection, bio-inspired computing and Web 3.0 are discussed. The section concludes with identification of research gap for the study.

2.1 Outlier Detection

Data science is an evolving field where data analytics is inevitable in the current scenario. With, increased inflow of data from various sources, decision making has become even more difficult [28]. Researches and practitioners have to now use data analytics for retrieving useful information from this data. The entire business analytics on the data can go in vain if this data is not analyzed correctly and at the correct time. Several data mining and content retrieval tools have come up for rescue to solve this huge data inflow and associated problems. Data outliers are data instances lying away from majority of the data points in that dataset. These points may occur because of several reasons including error while collecting the data, data variations or some deviations in the data itself that result into abnormalities. Figure 2.1 depicts a representation of outlier and normal data points.

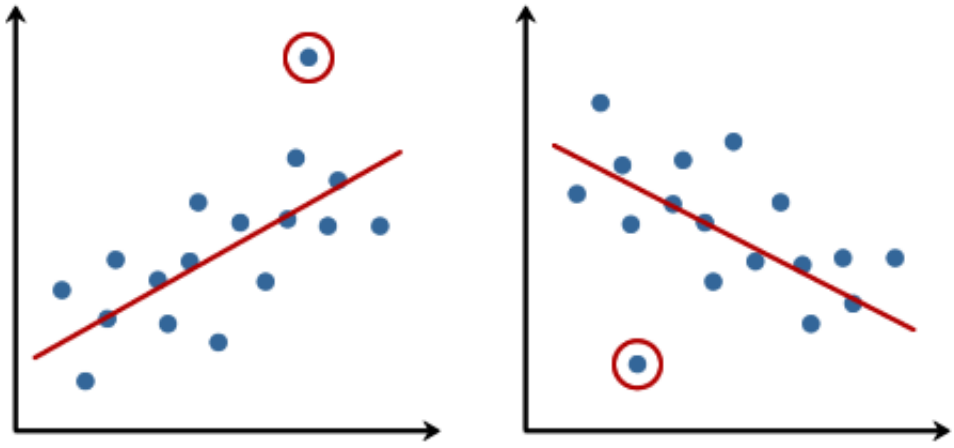


Figure 2.1: Diagrammatic Representation of Outliers

The approaches for outlier detection can be categorized into parametric and nonparametric methods, which refers to statistical and model-free respectively. The statistical approaches assume data distribution for the instances [29] [30] [31] or, are based on some sort of statistical estimate where the distribution for the data is not known [32]. Such approaches mark the data points that deviate significantly from the model as outliers. Such approaches come handy when working with high-dimensional data and wherever there is no advance information about the underlying statistical distribution of the dataset [33]. The category of non-parametric approach for detecting outliers is the one where the data-mining approach can be set apart. Such methods are also commonly known as distance-based approaches since they utilize the local distance measures and are often good for handling larger datasets.

Further, clustering based approaches comprise of the major chunk of data mining methods for detecting outliers, the approach considers small sized clusters of small sizes, as clustered outliers. The most popular approaches under this category include partitioning around medoids (PAM) and for large datasets the clustering large applications approach (CLARA) [34]. The CLARA was further extended for spatial outliers commonly referred to as CLARANS [35] and fractal dimension based approach [36]. Considering the clustering nature of these approaches and outlier detection being the secondary objective these algorithms are not usually optimized for detection of outliers. The criteria considered for outlier detection is often defined implicitly and cannot directly be inferred from approach [33].

Further, coming to the application domain, outlier detection is a widely explored area in various domains including wireless sensor networks [4], spam detection in social media [37], medical data [5] and fraud detection [6] amongst others. Literature highlights various outlier detection techniques and extensive reviews on methodologies [7] [8] [38] [39]. These approaches are categorized into nearest neighbor, statistical, clustering and classification based [9]. Methodologically, outlier detection has been explored in variety of different application domains. Existing reviews surrounding outlier detection techniques are based on different underlying statistics [4] [8] [7] [9] [38] [39]. Existing techniques are usually grouped into

statistical and machine learning based approaches [40]. A more detailed categorization groups these into information theoretic, statistical, classification based, spectral, clustering based and nearest neighbor [9].

There is extensive work in the domain of outlier detection using traditional heuristic approaches. Further, K-Means and KNN are simple yet efficient and popular approaches that can be used for mining outliers in clustering and classification datasets. The only drawback of these approaches is that they get stuck into locally optimum solutions. Further, in the current scenario, the explosion of information on the web and the richness of the content [41] have further made this analysis computationally extensive. With big data concepts coming in, the 3Vs of data, volume, variety and veracity increase the computational attributes when traditional machine learning algorithms are adopted on these data sets.

This has generated a need for using meta-heuristic approaches to expedite the analytics in such domains comprising of large amount of data [14]. However, most of the work is restricted to heuristic approaches however there are few recent evidences that have utilized meta-heuristic approaches by combining traditional machine learning models with existing optimization approaches. The existing literature discusses several meta-heuristic approaches specifically bio-inspired algorithms that may be utilized for solving analytical problems where data has volume and variety [42] [43].

2.2 Bio-inspired Computing

Bio-inspired optimization techniques have become a topic of great interest to researchers, since it has been noticed that nature has been evolving and there are certain species that have been in existence since centuries. It is amazing how these species have managed to survive for so long and results into applying their behavior and ways of living into computing. The most common meta heuristic approaches comprise of swarm intelligence algorithms that mimic the behavior of species in nature like bat [25], cuckoo [22] , firefly [24], wolf [26] and bees [23] to name a few. Meta heuristic approaches are being widely used in literature for various domains as

different as medical imaging [44] and cloud storage [45]. An extensive review on bio inspired algorithms and their applications highlight that these approaches have been successfully used in literature for similar analysis [14] [46] [47] [48]. Studies also highlight the use of optimization approaches specifically for cluster analysis [49] [50]. Figure 2.2 clearly depicts the evolution of heuristic, meta-heuristic and hyper-heuristic (combination of algorithms) algorithms. The most popular algorithms have been mentioned alongside the headings.

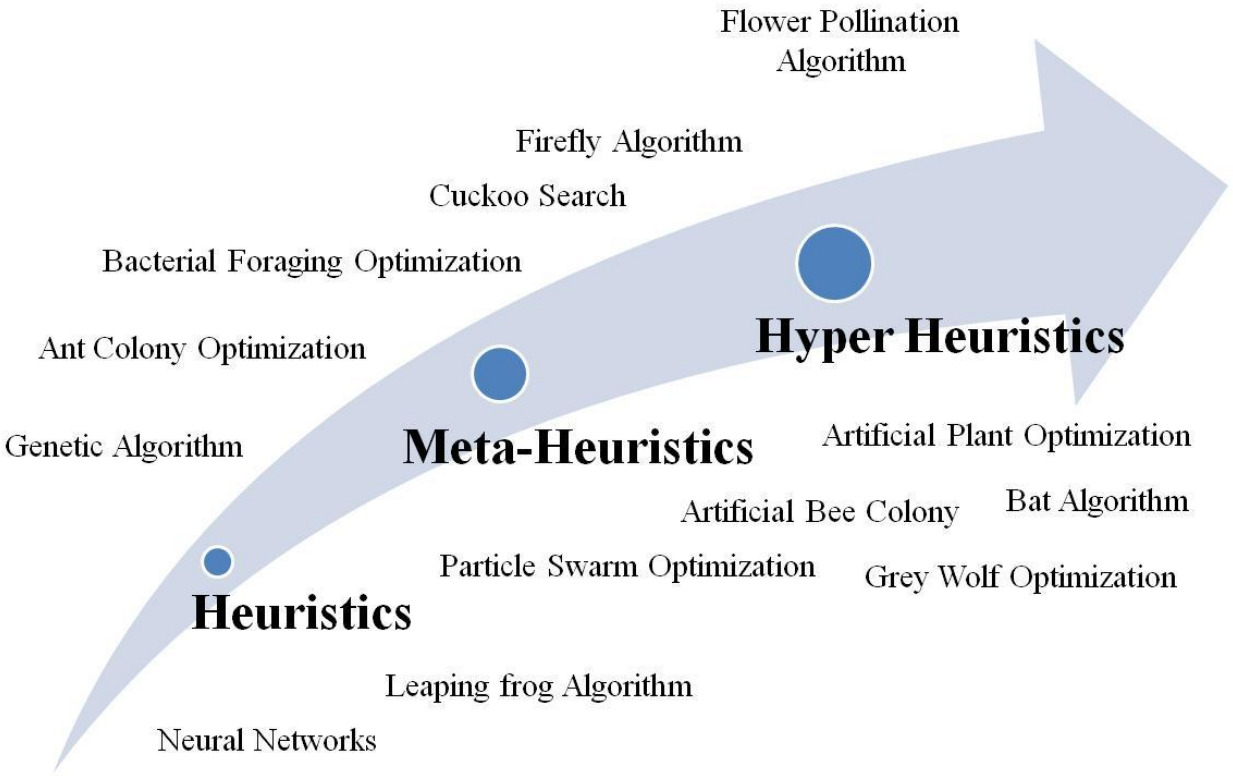


Figure 2.2: Evolution of Meta-Heuristic Approaches

(Adopted and recreated from Existing Literature [14])

Existing review studies in bio-inspired computing categorize the algorithms primarily into four set of quadrants that provides implications for researchers and practitioners. The quadrants divide the algorithms into theory development, applications, rediscovery and

commercialization. The first quadrant comprises of algorithms that showcase immense scope in terms of development of the algorithm. Literature still lacks evidences surrounding these algorithms and there is huge potential for improvement by introduction of improvements in terms of uncertainty, chaos and constraints. The latest algorithms in this quadrant comprise of wolf [51], grey wolf [26], lion [52] and amoeboid organism [53] amongst others that have been in literature for years.

The second quadrant targets the application domain since these algorithms possess the theoretical maturity however they can be explored and extended for different applications in variety of domains. The growth in this quadrant will have implications in diverse domains including intelligent systems, financial management, information systems and engineering to name a few. This quadrant comprises of bacterial foraging [21], bat [25], bee colony [23], cuckoo search [22], firefly [24] and flower pollination algorithm [54]. This work will have the primary focus on this quadrant by introduction of outlier detection using proposed hybrid versions of the same [14].

Further, the third quadrant as highlighted in the literature focuses on the rediscovery zone with leaping frog [18], shark search algorithm [55] and optimization using wasp colonies [56] coming into picture. The focus of this quadrant is on the algorithms that have been into the ecosystem for long but have failed to interest the researchers. This quadrant can again be rediscovered by introduction of fuzzy theory, chaotic optimizations and rough sets to gain researcher traction.

The last quadrant is the commercialization zone targeting algorithms that have been readily adopted by the research community. There has been plethora of studies surrounding these approaches focusing on different application domains. Due to ready adoption of these approaches even in industrial applications finding novelty surrounding the algorithms in this quadrant becomes an extremely challenging and constrained task. The quadrant comprises of genetic algorithm [16], neural networks [15], ant colony optimization [19] and particle swarm optimization [20]. It is now time to readily adopt the plethora variants available for these

approaches for commercialization. Figure 2.3 illustrates these quadrants for ready reference highlighting the main target algorithms for this approach [14].

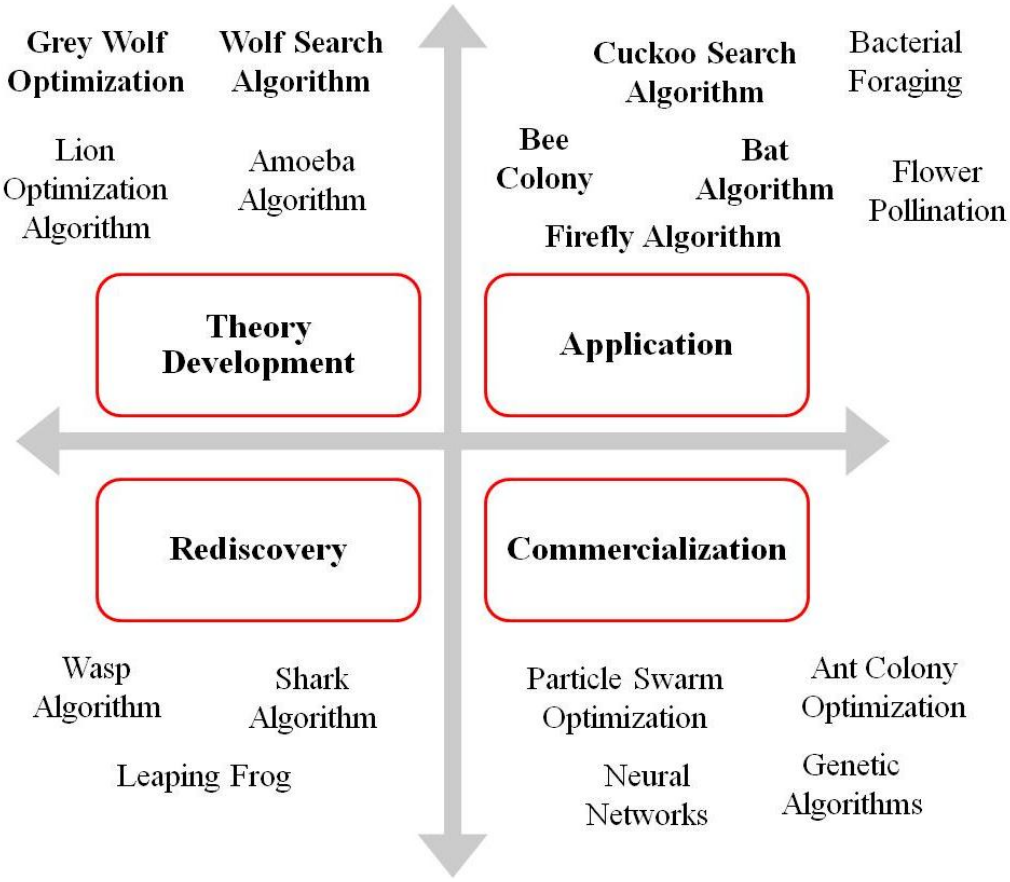


Figure 2.3: Quadrant Zones for Bio-inspired Algorithms

Recreated from review of bio-inspired algorithms [14]

Keeping in mind these constraints surrounding heuristic approaches for detecting outliers, meta-heuristics have gained immense popularity in this context. However, there are very few studies surrounding detection of outliers using meta-heuristics. Studies on this side primarily focus on detection of outliers using particle swarm optimization [57], firefly harmonic clustering [58], ant colony optimization [59], artificial neural networks [60], and genetic algorithms [61]. These studies discuss the growing application of meta-heuristics with a special focus on bio-inspired techniques for outlier detection in different domains of application. The bio-inspired search

space optimization methods avoid local optima while converging to the solution faster. Further, the application of such techniques in the trending domain of Web 3.0 is still missing. The subsequent-subsection highlights the importance of the domain.

2.3 Web 3.0 Domain

The increase in the use of internet has opened new avenues for several domains including e-commerce, e-governance, mobile applications, digital marketing, health care and politics to name a few [62] [63]. The patterns of internet use vary with the changing population and produces social capital [64]. Studies in the existing literature have thus analyzed the predictors of the use of internet and its subsequent drivers [65]. This has opened new doors for the virtual business environment and has completely changed the way business is done [66] [67]. Organizations are now adopting big data mining techniques to deal with this large amount of data for mining useful information [43].

Further, analytics surrounding the Web 3.0 domain including social media and websites has also taken leaps. It has been seen that over decades the use of information technology gives a competitive advantage to organizations. However, they need to be careful when utilizing this information for analytics as a lot of social media profiles under consideration may be potential spammers and may affect the sanctity of analysis. It further enables them to leverage the same for better consumer market understanding [68]. Studies also highlight that factors that affect the value creation in the e-business environments and have established frameworks for the same [69].

One of the most important part of Web 3.0 is social media. In recent times, organizations are also deploying resources to manage social media as it constitutes a substantial part for improving organic search results [70] [71]. This helps to direct potential customers to websites from search results and also from high integration with social media users [72]. There are frameworks defining functional building blocks namely identity, conversations, sharing, presence, relationships, reputation, and groups for expressing social media [73]. The world of

Web 3.0 and the huge influx of information make it infeasible to focus on all channels since business needs and channel receptivity depends on it.

Social media includes variety of platforms from online forums to blogs to discussion boards and social networking websites. With wide use of internet by the masses, the 21st century is witnessing an explosion of user generated content on the web [74]. This user generated content available on the social media networking platforms can be used in different fields like marketing, e-commerce, finances and so on by gauging the user's action and response to the events that occur. The information acceptability and content availability become major factors in influencing user behavior due to which social media has thus become a source of communication and engagement with stake-holders [75] [76] [77].

Further, there may be a difference of the perceived importance of channels of information and the subsequent trust on that information, based on the nature of information and who propagates it [78]. So organizations have started developing strategies for social media platforms such as YouTube, Facebook, and Twitter and devise mechanisms to leverage social media in reaching an important audience of users by modeling popularity [79] [80] . Literature also investigates social media's emerging importance in the current market research as it enables sharing and discussing of information with others having similar interests related to politics [81], technology and business to name a few [82] [83] . Literature also talks about the business value of social media and provides an input to social media selection process [84].

Another important aspect of the interactive Web is the websites and influencer blogs available in the ecosystem. With the advent of Web 3.0 and increased use of the internet by people, the visibility of content on the web is of prime importance. Literature highlights the important factors for online marketing and their relevance in making the content popular among the users [85]. Now, a lot of traffic to these organization websites comes from organic search queries in search engines like Google [86], Bing and Yahoo to name a few.

Search engines prove to be a great tool in quest to locate online information by people. People may have varying browsing behaviors [87] [88]. About 92% of adults in the U.S. make use of search engines to garner online information that they require [89]. There are varying characteristics and changes in web searching for different search engines and case studies [88]. Organizations have thus started using search engine marketing (SEM) and adopting various strategies to promote their content on the web [90]. Literature highlights the use of SEM in various domains including tourism, e-commerce and marketing [91] [92] [93]. It is noticed that people generally tend to use the top results from their search query, making the rank of web pages of vital importance for the companies.

As a result, organizations have started adopting strategies for brand positioning using SEM [94]. An important technique used by companies to improve the ranking of their pages is search engine optimization (SEO) a major tool in SEM that plays a critical role in increasing the web pages' visitor count. This is usually done by ranking it higher on the search results often using keywords that describe the website's content [95]. SEO techniques are thus critical in increasing the visibility of a website on the internet and attract greater organic search traffic [96]. SEO techniques may be categorized into two: White Hat and Black Hat SEO [97]; White Hat SEO techniques are within the SEO guidelines, deemed legal and ethical, usually comprises of using quality content, titles, meta data, keyword research, effective use of keywords and inbound links. On the contrary, black hat SEO techniques do not lie within the SEO guidelines and often result in poor ranking and blacklisting of the website from the search engine when detected. These techniques comprise of cloaking, use of doorway pages and invisible elements [98].

It is evident that individuals and organizations both are in a constant race for visibility on the web including search engines, blog and social media platforms. Both profit and non-profit firms nowadays are majorly interested in online marketing of their ideas, services, products and projects. These organizations have realized that the web plays an inevitable role and attracts significant marketing opportunities. Considering these evidences from the literature, it becomes critically important to identify aberrations in the Web 3.0 domain including social media and

search engines. The current work thus attempts identify outliers (distinctly different data points) of various forms in the domain of Web 3.0 using bio inspired computing techniques.

2.4 Conclusion

Bio-inspired computing algorithms are known for optimizing an objective function either for minimization or maximization. These meta-heuristic approaches when integrated with traditional machine learning algorithms help in avoiding locally optimal solutions expedite the convergence to the solution and enhance the accuracy of the obtained solution. In addition to the methodological contribution, the work addresses concerns in the domain of Web 3.0 surrounding outliers. The identified research gaps can thus be summarized as following:

- Lack of studies surrounding the use of hybrid bio inspired computing techniques for outlier detection.
- Limited evidence of application of outlier detection in the domain of Web 3.0.
- Very few studies in the domain of Web 3.0 using bio inspired computing algorithms.

To address the concerns surrounding detection/identification of outliers in the Web 3.0 domain the study proposes three research objectives with sub-problems having practical implications in the industry. The research objectives attempt to address practical problems faced by organizations and individuals with the advent of Web 3.0 and increased use of internet in every facet of communication and engagement. The findings of this work have policy implications in domains of e-commerce, social media and influencer marketing. The objectives include outlier detection in supervised domain, unsupervised domain and integration chaos theory for outlier detection. The proposed research methodology proposes hybrid bio-inspired computing algorithms to solve the identified concerns surrounding identification of outliers in the Web 3.0 domain.

CHAPTER 3

3. DATA DESCRIPTION

This chapter focuses on the description of datasets used in the work. As highlighted in the research problems the work targets three research problems surrounding outlier detection in supervised scenario, unsupervised scenario and integration of chaos for outlier detection. Each research problem targets two sub-problems and the datasets used for the same are described in the current section.

3.1 Publically Available Datasets

The first problem surrounding outlier detection for supervised scenario utilizes two publically available datasets from UCI repository. The first dataset is classification dataset called Iris is often referred to as one of the best datasets available in the pattern recognition domain. The dataset comprises of 3 output classes, each having 50 instances. Each output class refers to a variety of plant. Statistically, one output class is linearly separable from the remaining two while the two are not from each other. The dataset comprises of 4 attributes namely sepal length, sepal width, petal length and petal width are measured in centimeters. These attributes are modeled together to predict the type of the plant which can be Iris Setosa, Iris Versicolour and Iris Virginica. The statistical estimates for the dataset are provided in Table 3.1.

Table 3.1: Statistical Estimates for Iris Dataset

	Minimum	Maximum	Mean	SD	Class Correlation
Sepal Length	4.3	7.9	5.84	0.83	0.7826
Sepal Width	2.0	4.4	3.05	0.43	-0.4194
Petal Length	1.0	6.9	3.76	1.76	0.9490 (High)
Petal Width	0.1	2.5	1.20	0.76	0.9565 (High)

The second dataset considered for the work is the Abalone which is based on a biological species usually found in Tasmania. This regression dataset is used for predicting the abalone age using physical metrics. The process of finding out the age is by performing the cut in the shell through its cone, staining the shell and subsequently the number of rings seen are counted using a microscope. This process is often very taxing and challenging and thus there are several physical attributes that can be used to predict the age. The dataset comprises of 4177 instances to predict the age. The attributes include the gender, the length, height, diameter, weight (Whole, Shucked, Viscera and Shell) and the number of rings, the same are described in Table 3.2

Table 3.2: Description of Abalone Metrics

Attribute Name	Data Type	Measurement	Description
Gender	Nominal	M, F, and I (infant)	
Length	Continuous	mm	Longest Shell Measurement
Diameter	Continuous	mm	Perpendicular to Length
Height	Continuous	mm	With meat in shell
Whole Weight	Continuous	grams	Whole Abalone
Shucked Weight	Continuous	grams	Weight of meat
Viscera Weight	Continuous	grams	Gut Weight (after bleeding)
Shell Weight	Continuous	grams	After being dried
Rings	Integer		+1.5 gives the age in years

The basic statistical estimates for the numeric data like maximum value, minimum value, mean, Standard Deviation (SD) and correlation for the dataset are depicted in Table 3.3. These datasets are used to validate the proposed approach for outlier detection.

Table 3.3: Statistical Estimates for Abalone Dataset

	Length	Diameter	Height	Whole	Shucked	Viscera	Shell	Rings
Minimum	0.075	0.055	0.000	0.002	0.001	0.001	0.002	1
Maximum	0.815	0.650	1.130	2.826	1.488	0.760	1.005	29
Mean	0.524	0.408	0.140	0.829	0.359	0.181	0.239	9.934
SD	0.120	0.099	0.042	0.490	0.222	0.110	0.139	3.224
Correlation	0.557	0.575	0.557	0.540	0.421	0.504	0.628	1.0

3.2 Influencer Website Blogs

The second problem surrounding supervised scenario adopts mixed research methodology where in the data collected surrounding the website KPIs for 2751 influencer blogs on unique domains. A statistical t-test is conducted on the normalized data for the two sets of influencer web domains, with low and high spam score.

The data is extracted through an API from the SEO Rank website that provides a holistic list of selected metrics provided by various data providers like Majestic, Ahref, Moz, SemRush and Webmaster tools. These data providers have developed ranking mechanisms that are used worldwide for identifying the position of a page in organic search. Table 3.4 provides the holistic list of metrics extracted.

Table 3.4: List of metrics used for identification of Influencer Blogs

	Data Provider	**Metric**	**Description**
1.	Moz	Domain Authority	Prediction of the ranking of domain on search engines. Depends on links, Moz Rank and other metrics.
2.		Page Authority	Prediction of how a given URL may be ranked on search engines, associated with number of links, Moz Rank, and others.
3.		Moz Rank	Link popularity score indicative of importance of the

			page on the web.
4.		Moz Trust	Link trust checks for links from trustworthy sources.
5.		Links In	Links to the web page, includes equity, or non-equity both internal and external links.
6.		External Equity Links	Number of external equity links to the URL
7.		Spam Score	Based on number of sites penalized (de-listed) containing links to the web page.
8.	Alexa	Alexa Rank	Global Alexa rank of webpage
9.		Alexa Links number	Number of links to the web page
10.		Country Rank	Alexa Rank in the popular country.
11.	Majestic SEO	Citation Flow	Uses site link counts to the web page to see how influential the page is.
12.		Trust Flow	Trustworthiness of the page based on link to trustworthy neighbors.
13.		External Back Links	Total external back links to the web page
14.		Referred Domains	Total unique domains having links to the website
15.	SemRush	SemRush Rank	Domain rank by SemRush
16.		URL Links	Links to the mentioned web page
17.		Hostname Links	Links to the domain
18.	Ahrefs	URL Rating	Strength of web pages' back link profile and its chances of being ranked high in Google.
19.		Domain Rating	Strength of website's domain back link profile.
20.		Ahrefs Rank	Ranking based on size and quality back link profile
21.		Live & Fresh Index	List of live and dead links for the website

3.3 Twitter Buzz Instances

The data used for first sub-problem in unsupervised scenario is a Twitter data with metrics describing the discussions held on the platform. The data is publicly available on UCI repository for conducting supervised regression/classification analysis [144]. The current work uses the same for identifying outlier instances in the form of buzz. The dataset includes 11 attributes that are modeled to compute the mean count of active discussions which is indicative of the popularity of the topic instance. This attribute is further used to classify the instance into buzz and non-buzz based on the designated threshold component.

The dataset is a projection of the value of the 11 attributes at 7 different time instances covering the observations for a specific topic over a week. This results in a total of 77 attributes for all the discussions that are available for prediction of the active discussions in the temporal space. Table 3.5 describes the various attributes.

Table 3.5: Attributes used for identifying Twitter Buzz

Attribute	Category	Attribute Name	Description
A1	Network Level	Number of Created Discussions	Measure of the number of discussions created within a specific time (e.g. 1 hour of the tweet getting posted) on a tweet's topic.
A2		Author Increase	Number of new contributors who react to that tweet within a specific time (e.g. 1 hour of the tweet getting posted).
A3		Attention Paid	Measures the attention/number of views that the tweet gets on social media within a unit time. It is affected by the number of interacting contributors.
A4		Burstiness Level	Measure of the ratio of created discussions to the mean length of each discussion in terms of a unit of time.
A5	Tweet Specific	Atomic Containers Tweet	Measure of number of containers/topics of discussion surrounding the individual tweet.
A6		Attention Level Tweet	Measures attention that is paid to the tweet on its creation based on the topic of the tweet.
A7	Topic Level	Contribution Sparseness	Measure of contribution spread over discussions for the particular topic over a period of time.
A8		Author Interaction	Measure of the average number of authors that interact on the topic within discussions created by individual tweets.
A9		Number of Authors	Measure of the number of authors interacting within a specific time period (e.g. 1 hour of the tweet getting posted) on a specific topic.
A10		Mean Discussions Length in time	Measure of the mean discussion length in time that is incited by a topic from all the contributors of the topic.
A11		Average Number of Discussions	Measure of the average number of discussions that involve the topic within a specified time.

This work uses the average value for each attribute across the week's instances for the analysis. This results in the final dataset comprising of 583,249 instances which is modeled using 11 independent attributes. The final output attribute is the number of active discussions depicting the popularity of the topic instance.

3.4 Twitter Fake Profiles

The second problem for the unsupervised scenario focuses on detection of fake profiles. Twitter data is collected using standard APIs and is usually unstructured in nature with a lot of emoticons and informal expression of text. The data collected is also more enriched in a way that it contains HTML links, hashtags and @mentions. Thus, the analysis of this data collected from social media sites becomes more challenging than the data available in standardized databases. The approach involves use diverse analytical methodologies and approaches for pre-processing of data to achieve the metrics which may be finally needed for mining useful information.

This study extracts the data using R's 'TwitteR' API to fetch user profile and raw tweet data on periodic basis. The analysis was done using a total of 5,55,684 tweets. For understanding which metrics could be useful for explaining user centric behavior in Twitter, first through a review of literature on content virality and information diffusion [130] [137] [138], impact of social media [129] parameters were identified. Table 3.6 gives a list of the 27 parameters that were taken into consideration, these have been adopted from existing studies [129] and comprise of content including sentiment mining, descriptive and network analytics parameters.

Table 3.6: List of 27 metrics for Fake Profile Identification

@ Mentions	Unique Hashtags	Polarity Stability
Centrality	Activity and Visibility	Lexical Diversity
Follower Count	Betweenness	Term Adjacency
Following Count	Word Frequency	Sentiment analysis
Hash tag frequency	Tweet Frequency	Link diversity
HTML link count	Response to @ mentions	Emotion Stability
Network density	Added to lists	Clique
Replies	Retweets per tweet	Propinquity
Words per tweet	Favorite count	Reciprocity/Mutuality

Table 3.7: Description of final set of metrics for Detection of Fake Profiles.

Metric Name	Metric Description
Emotion Stability (M1)	A metric that measures the deviation in emotional score for the given emotions-joy, sadness, surprise, disgust, fear and anger for a user in the last 50 tweets.
Polarity Stability (M2)	A metric that measures the deviation in polarity score for both negative and positive polarity for a user in the last 50 tweets.
Hashtag Frequency (M3)	A metric that measures average number of hashtags in the last 50 tweets.
Unique Hashtags (M4)	A metric that count the number of uniquely used hashtags in 50 recent tweets.
HTML Link Count (M5)	A metric that measures average number of HTML links in each status.
Unique Words (M6)	A metric that measures average number of unique words in each status.
@ Mentions (M7)	A metric that counts @ mentions of other users in 50 recent tweets.
Lexical Diversity (M8)	A metric that defines the fraction of unique words upon the total words used by a user in 50 recent tweets.
Status Count (M9)	A count of statuses updated by the user.
Favorites Count (M10)	A count of number of tweets a user has liked.
Friends Count (M11)	A count of number of users the user follows.
Follower Count (M12)	A count of number of followers of the user.

3.5 Mashable News Content

The last research problem surrounding the integration of chaos for outlier detection includes a sub-problem for identification of popular content. The dataset used for the analysis comprises of 39,797 articles from a popular content publishing website called Mashable. The data is publically available on UCI repository. Each article is expressed by a set of 58 predictive parameters that includes descriptive metrics pertaining to links, words, digital media (images and videos) and publishing time. It further comprises of metrics surrounding sentiment and polarity of keywords, topics and several metrics that are extracted using advanced text mining techniques. The articles focus on primarily six domains including "Tech", "Entertainment", "Social Media", "Lifestyle", "World" and "Business".

Since the values for the metrics belong to varied ranges, the data was normalized to 0-1 using a min max normalization approach. Further, based on the shares threshold the data is divided into popular and not popular and a statistical t-test is conducted to identify significant metrics. The Table 3.8 describes the significant metrics used for the purpose of analysis in the current study. It comprises of 19 independent metrics and one outcome metric which is used to predict content popularity. The study considers a 95% confidence interval and the metrics with a p-value less than 0.05 have been finally considered for analysis.

Table 3.8: Description of Metrics for Online Content Popularity

M. No.	**Description of metrics for online popularity**	
	Metric	**Description**
M1	Link count	The count of links embedded in the content
M2	Mashable link count	The number of links referring to Mashable articles
M3	Image count	Count of embedded images
M4	Video count	Count of embedded videos
M5	Keyword Count	Keyword count in the article
M6	Mashable article share	Number of times the article is shared in Mashable

M. No.	Description of metrics for online popularity	
	Metric	**Description**
M7	Closeness to topic 1	Closeness of article to “Tech” as obtained using topic modeling
M8	Closeness to topic 2	Closeness of article to “Entertainment” as obtained using topic modeling
M9	Closeness to topic 3	Closeness of article to “Social Media” as obtained using topic modeling
M10	Closeness to topic 4	Closeness of article to “Lifestyle” as obtained using topic modeling
M11	Closeness to topic 5	Closeness of article to “Business” as obtained using topic modeling
M12	Content subjectivity	Subjectivity of the article text
M13	Content sentiment	Sentiment of the article text
M14	Avg. positive polarity	Average positive polarity of the article
M15	Avg. negative polarity	Average negative polarity of the article
M16	Title subjectivity	Subjectivity of the article title
M17	Title polarity	Polarity of the article title
M18	Absolute subjectivity	Level of absolute subjectivity
M19	Absolute polarity	Level of absolute polarity
M20	Total shares	Target output variable for total number of shares

3.6 Search Engine Marketing Websites

The second sub-problem is surrounding mining outlier websites on search engines is of prime importance to have a comparative analysis of the same. As discussed there are several metrics for computing a website’s rank. These metrics have been given by different companies and comprise of over lapping criteria that suit their individual needs. For the purpose of this study we are considering two separate cases for different organizations. The dataset used for the

analysis of this study thus comprised of two different cases: Case 1 and Case 2 with 1070 data points and 1682 data points respectively with each data point belonged to a separate website. Since the number of metrics is large and computationally complex, a mix research methodology is this adopted that includes a Delphi Study for finalization of metrics followed by bio inspired algorithms to mine outliers. Also Delphi is needed since organizations choose the different sets of metrics for selecting suitable websites for partnering in digital marketing campaigns as objectives of campaigns differ.

For the identification of relevant parameters for our study, a Delphi study was conducted. Delphi is a quantitative technique and one of the alternatives to achieve consensus for a problem. It seeks opinions from the group of experts in an iterative manner with rounds of questions answered. A summary of responses is generated and redistributed among the panelists for discussions in the subsequent rounds. For the purpose of this study, Delphi is used to check the applicability of the list of 16 metrics for the identification of final metrics to be considered for analyzing outlier websites. A list of metrics that are frequently used by different providers for analyzing the rank of web pages is described in Table 3.9.

Table 3.9: Description of Metrics for Web Page Ranking

S. No.	Metric	Description
1.	Page Rank	Metric used by Google Search to rank websites in their search engine results.
2.	Page Authority	Moz Page Authority represents the authority of a specific individual pages or URLs on a 1 to 100 scale. Page Authority is a compilation of several Moz metrics.
3.	Domain Authority	Moz Domain Authority It represents the authority of the domain on a 1 to 100 scale. It is a measure of the predictive ranking strength for the domain including the sub-domains.
4.	MozRank	MozRank represents the global link popularity on a 1 to 10 scale.

5.	MozTrust	Represents Moz's global link trust score. It measures link trust, it checks for trustworthy sources (e.g. Government, university websites) for receiving links.
6.	Citation Flow	A metric given by Majestic SEO that uses the count of sites linked to it to predict how influential a web page is. The more links the site has, the higher the CF will be.
7.	Trust Flow	Another metric by Majestic SEO to predict trustworthiness of a web page. Back link's nearness to the trusted and aged domains is used to compute the trust flow.
8.	Alexa Rank	Alexa Rank refers to the calculation of Alexa traffic ranking using user reach and page views.
9.	URL Rating	It is a metric given by Ahref that measures the strength of back link profile or a target web page, Often measured on a logarithmic scale from 1 to 100.
10.	Domain Rating	A metric by Ahref to compute overall back link profile of a given webpage on a logarithmic scale from 1 to 100.
11.	Ahref Rank	Ahref's Rank compares the back link of the website. It uses the size and domain rating for computation.
12.	Back Links	Back links to a webpage are its incoming links when it links to another web page. These have been used as a metric to rank pages in the past.
13.	Total Links	Represents the total number of links to a web page. It also includes, External/Internal Links and Do-follow/no-follow links.
14.	Google Index	The count of specific web pages that Google is able to crawl or index on a website.
15.	Social shares	The shares for the web page on social media platforms like Facebook, Twitter, Google Plus and others.
16.	Domain Age	It is metric that gives the approximate age of a website and is often a criterion for ranking web pages on search engines.

CHAPTER 4

4. OUTLIER DETECTION IN SUPERVISED SCENARIO

This chapter explores the use of hybrid bio-inspired computing algorithms for detecting outliers in the supervised scenario. The supervised scenario may comprise of classification or regression datasets wherein we have a continuous or nominal output label. The motivation behind these studies is to avoid locally optimum solutions and minimize convergence iterations which are significantly high when it comes to traditional machine learning approaches implemented for big datasets. The sections implements hybrid grey wolf optimizer, wolf search algorithm and bat algorithm. The latter have been explored for detecting outliers among influencer blogs.

4.1 Proposing Hybrid Algorithm for Outlier Detection

4.1.1 Introduction

Traditional machine learning algorithms specifically k-nearest neighbors have been a popular choice when it comes to outlier detection. It is a simple yet effective approach to adopt. The only fallback is that such approaches tend to get stuck local optima and often take a lot of time to converge for large datasets. With the exponential increase in the amount data being generated from different sources there has been a critically essential need to mine important and useful information from it. Data mining is becoming an integral part of any business. However, with the amount of data under consideration such approaches are becoming challenging when used in the real life scenarios with different data distributions and quantity of data being dealt with.

Meta-heuristics on the other hand specifically bio-inspired algorithms are known for optimizing an objective function and converging to a globally optimum solution with faster convergence. Outlier detection approaches have been popular but have rarely been explored through this lens of bio-inspired computing. The proposed approach uses a combination of a popular bio-inspired optimization based on the hunting behavior of wolves [10] and nearest neighbor approach.

When considering optimization algorithm integration, it becomes essential to identify an objective function along with the motive of using it. The problem can either be a minimization or a maximization problem.

4.1.2 Background

Outlier detection is a widely explored area in various domains including wireless sensor networks [4], spam detection in social media [11], medical data [5] and fraud detection [6] amongst others. Literature has several existing reviews on outlier detection techniques in various domains [7] [8] [12] [13]. KNN is one of the first choices for detecting of outliers. However, being a heuristic approach it often falls into local optima. In the domain of analytics, the best solution for an objective function in a minimization or maximum solution can get stuck into local optima. The Figure 4.1 diagrammatically explains the local and globally optimum solutions for the objective function *f(x)*.

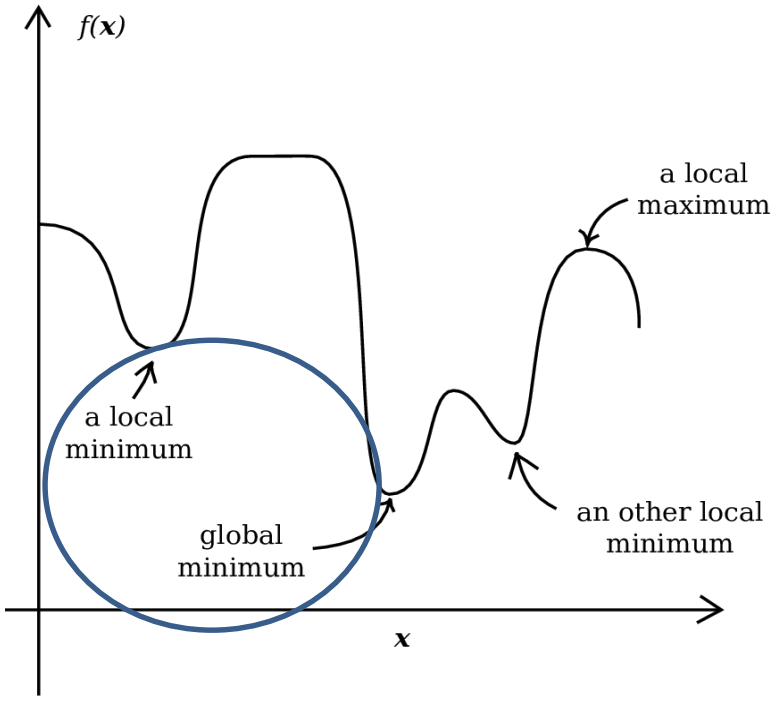


Figure 4.1: Illustration of Local Optima and Global Optima

Further, due to increased variety and volume of the data, it has become computationally complex to analyze data using simple machine learning approaches. Thus, with these existing issues in heuristic approaches, meta-heuristic approaches have become prominent over the decades [16] [17]. The most common meta-heuristic approaches comprise of swarm intelligence algorithms that mimic the behavior of species in nature. Literature provides extensive reviews for bio-inspired algorithms and their applications in various domains [24] [25] [26] [27].

However, literature showcases limited evidences of usage of these meta-heuristic approaches in the domain of outlier detection. There are very few studies that have proposed an outlier detection approach using meta-heuristics [30] [31]. Further, these studies do not clearly cover the aspect of convergence while reaching to the final solution and have limited scope for small datasets.

4.1.3 Methodology

This work utilizes the Grey Wolf Optimizer (GWO) which works surrounding the hunting behavior of grey wolves [26]. In the current problem discussion for outlier detection using the integrated meta-heuristic approach, the accuracy of KNN becomes the objective function. The work uses a modified GWO with the position of the top three search agents (X_α, X_β, and X_δ) not being updated with each iteration. The α being the best position for the wolves is obtained as the output of KNN, β, δ and ω the next three positions respectively. Figure 4.2 depicts the pseudo-code for the proposed approach.

Pseudo-code for Modified GWO using kNN Algorithm

```
Begin
Initialize the grey wolf population
Initialize Test Data as Number of grey wolves (NumGW)
for each grey wolf NumGW
      /*calculate Accuracy and Position of the grey wolf */
/* i=1 Iteration 1 */
      R ← random number (between 1 to NumGW) is generated
      R is checked from a database of numbers so that R should not repeat
      /* Take data points one by one from TestData as position*/
Position ← TestData(R)
      Pass Position to kNN as input and get kNN Output
      Accuracy ← Target Value – kNNOutput
      Best_Value ← Acuuracy
      Best_Position ← Position
/* i=2 Iteration 2 */
for each grey wolf NumGW
      if (Best_Value > Acuuracy)
      then
            Best_Value=Accuracy
            Best_Position=Position
      End
End
```

Figure 4.2: Pseudo-code for modified GWO using KNN algorithm

The fitness of X_α (best search agent) is computed over iterations through the above mentioned approach. The best score and best positions for best ‘n’ independent variables are computed with ‘n’ being the number of attributes under consideration in our dataset. The primary reason for doing so is the α position which is being pre-calculated using the KNN by optimizing the accuracy making the β, δ and ω solutions insignificant in the current context.

Further, since KNN is used to compute the best set of wolf positions for detecting outliers. A data point having 'n' dimensions, basically each row comprising of ‘n’ columns, as the (Target-Output) comprising of the minimum value being obtained. This makes it a minimization problem where the objective function works towards minimizing the value of (Target-Output)/KNN accuracy. The best values obtained as output from the KNN act as the best (α)

positions for the GWO part of the approach where the wolves hunt the prey or in other words where KNN achieves the best accuracy. The details of the GWO-KNN approach in the form of a flowchart are described in Figure 4.3.

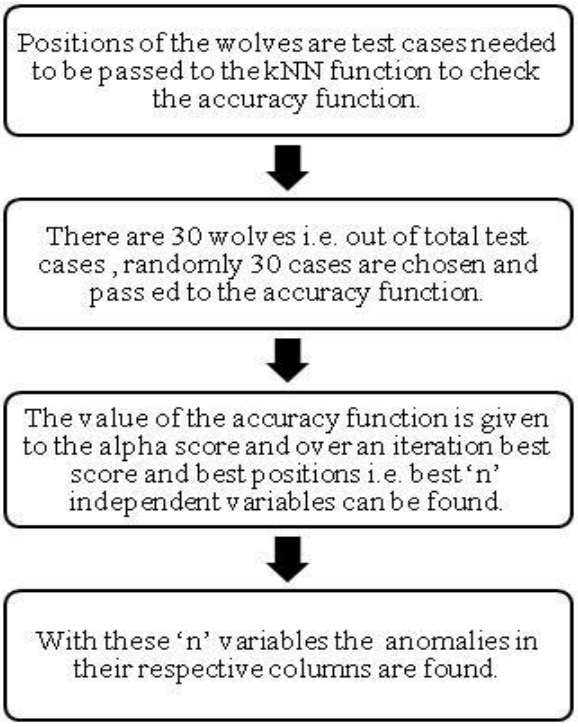


Figure 4.3: Flowchart for Proposed GWO-KNN Algorithm

4.1.4 Analysis and Findings

The proposed approach is tested by on publically available classification dataset Iris. The dataset is available on the UCI Repository for machine learning experiments. The details of the same are expressed in Section 3.1. For the purpose of this algorithm we consider the instances of the minority class as anomalies. This was just for the purpose of validating the results since for classification of one class the other two classes seemed to be significantly different. The proposed approach produced correct output for this dataset and can prove to be beneficial for other supervised datasets by tuning the value of 'k' for the underlying KNN algorithm to which

the grey wolf is integrated. The Iris datasets comprises of 4 attributes that help in identifying the class of the plant which could be Iris Setosa, Versicolour or Virginica.

For the purpose of experimentation, 50 instances from the class Iris Setosa along with 30 from the Iris Versicolour are obtained making the Veriscolour class as the minority with respect to Setosa instances. As described in the proposed approach, the minority class instances should appear anomalous when compared to the one in majority and should lie farther way forming a different cluster. The analysis proved as per assumption and the individual plots for the 4 Iris dataset attributes including sepal length and width, petal length and width are illustrated individually in the Figure 4.4, Figure 4.5, Figure 4.6 and Figure 4.7. The instances marked in red are the outlier instances while the ones in blue are the non-outlier instances.

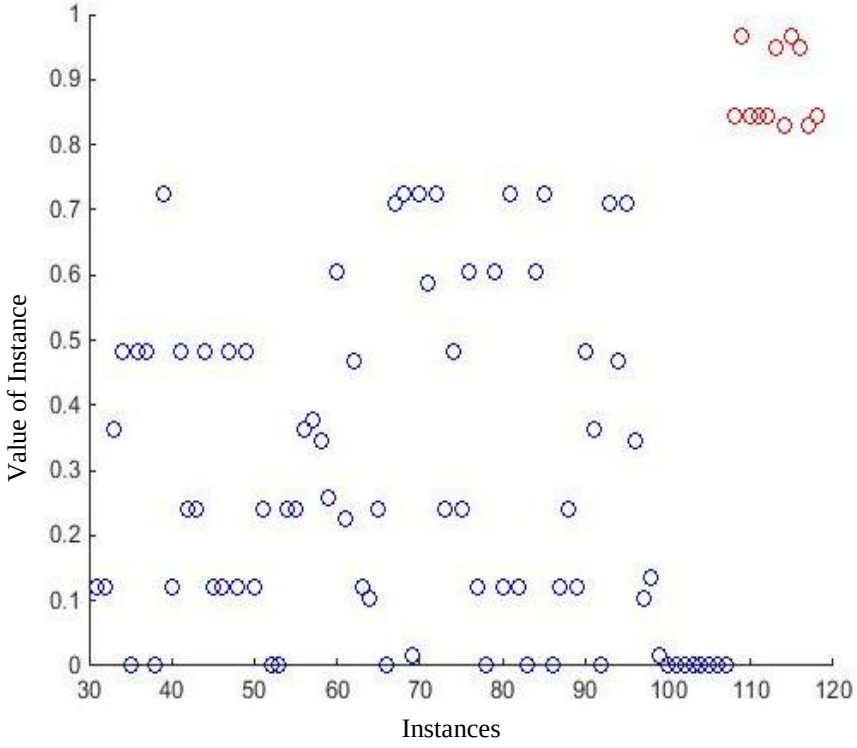


Figure 4.4: Outlier Plots for Iris (Sepal Length)

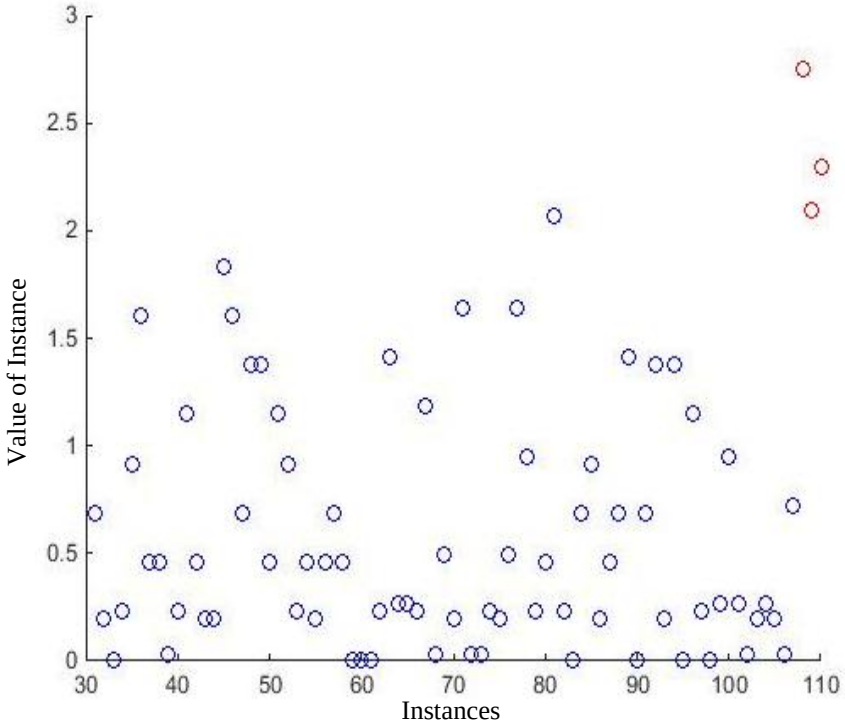


Figure 4.5: Outlier Plots for Iris (Sepal Width)

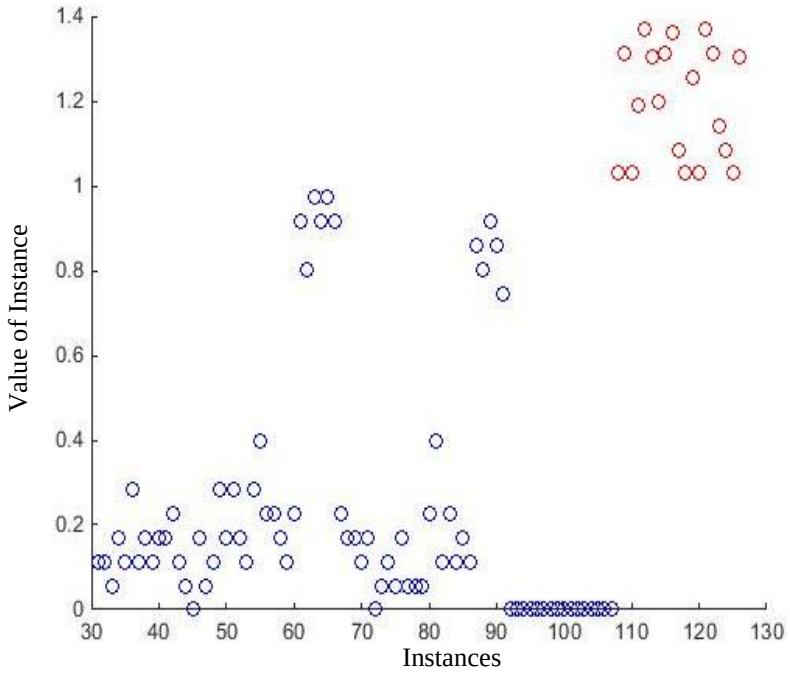


Figure 4.6: Outlier Plots for Iris (Petal Length)

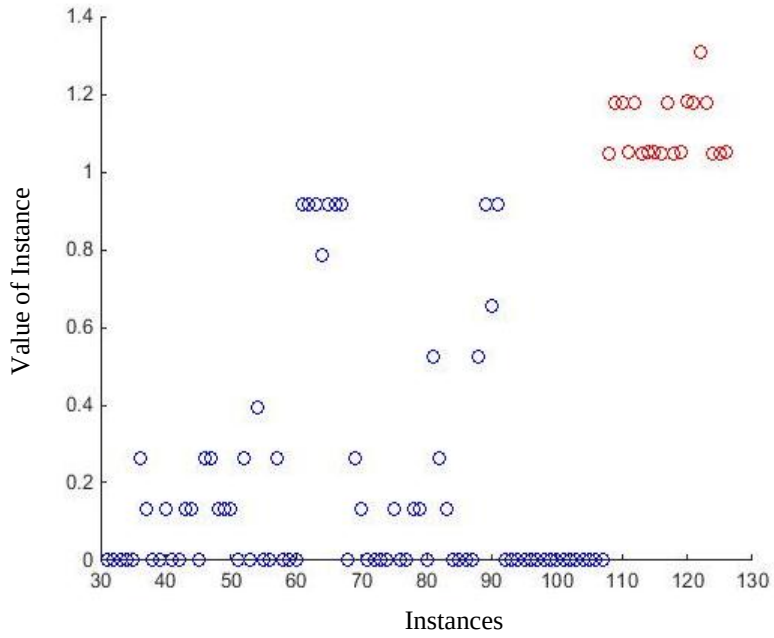


Figure 4.7: Outlier Plots for Iris (Petal Width)

Similarly, for the regression dataset Abalone which is again a supervised regression dataset available on the UCI repository, comprising of more than 4000 instances with 8 attributes being modeled to predict the final outcome, the algorithm produce 8 separate outlier plots for each of the attribute. Figure 4.8, Figure 4.9, Figure 4.10, Figure 4.11, Figure 4.12, Figure 4.13, Figure 4.14 and Figure 4.15 depicts the results for the same. Each plot illustrates the outliers for that attribute in red and the non-outlier data points in blue.

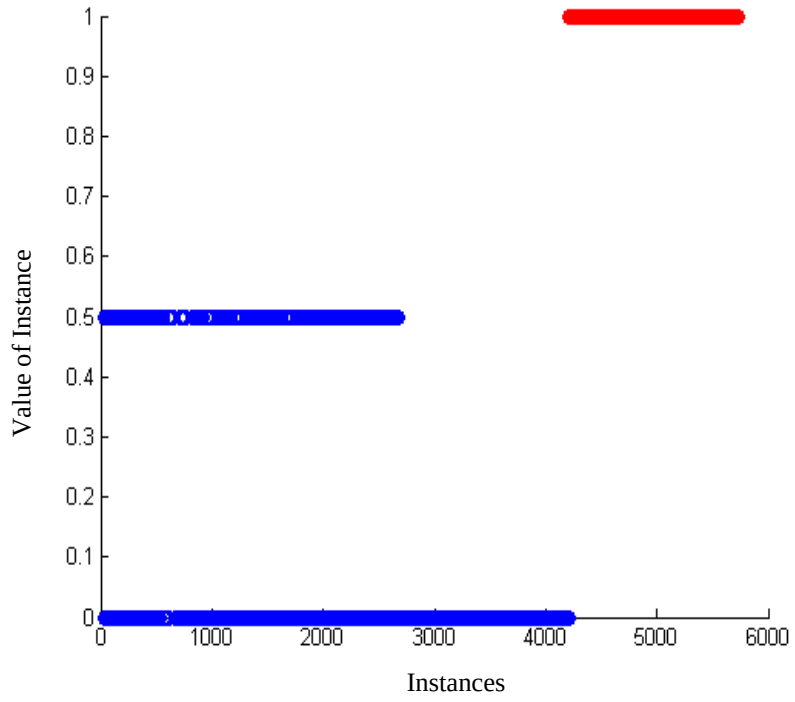


Figure 4.8: Outlier Plots for Abalone (Gender)

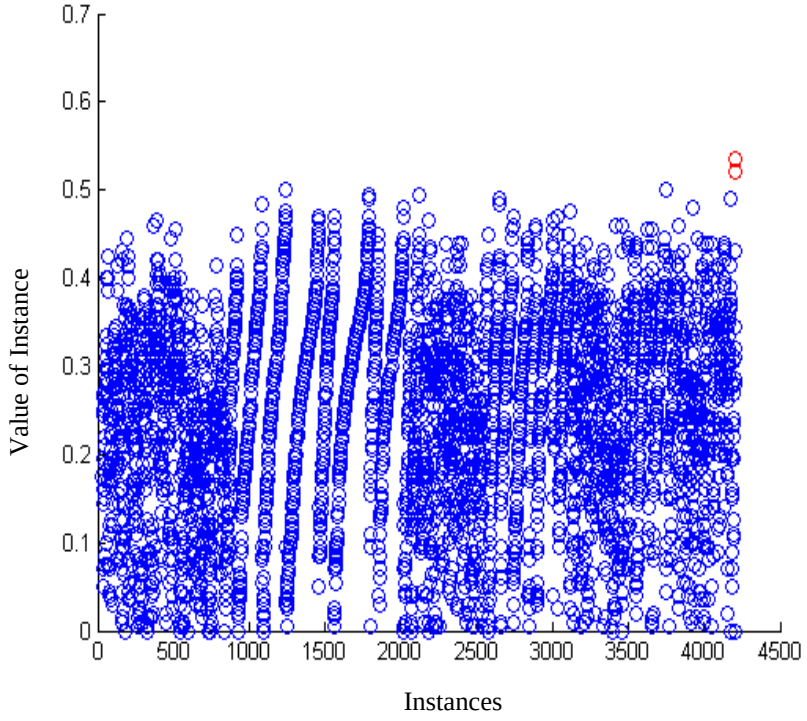


Figure 4.9: Outlier Plots for Abalone (Length)

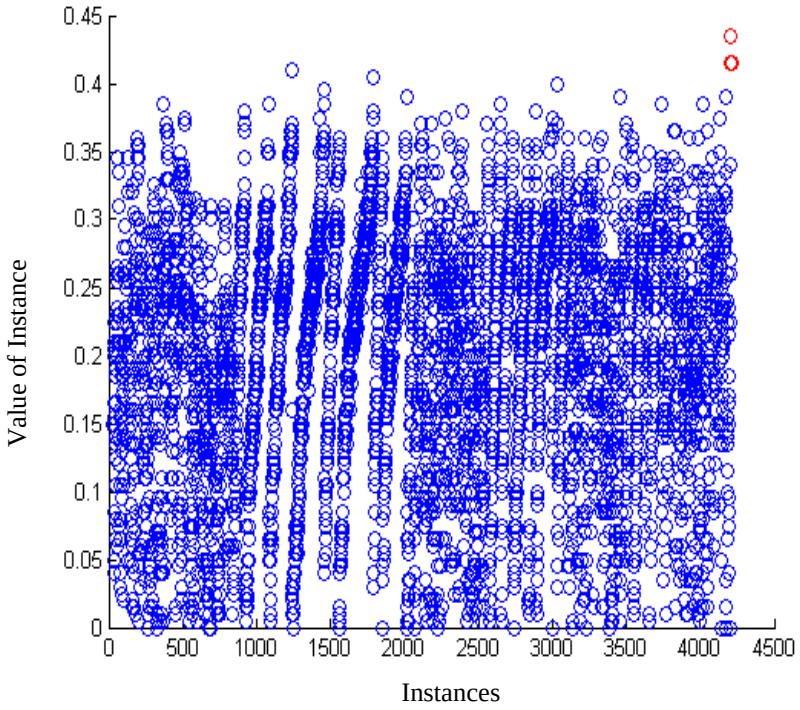


Figure 4.10: Outlier Plots for Abalone (Diameter)

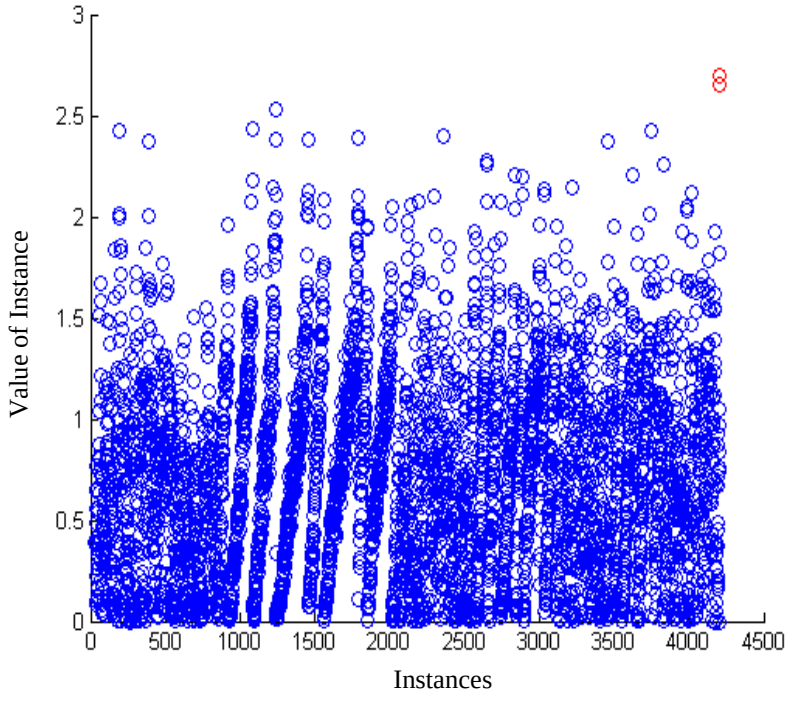


Figure 4.11: Outlier Plots for Abalone (Height)

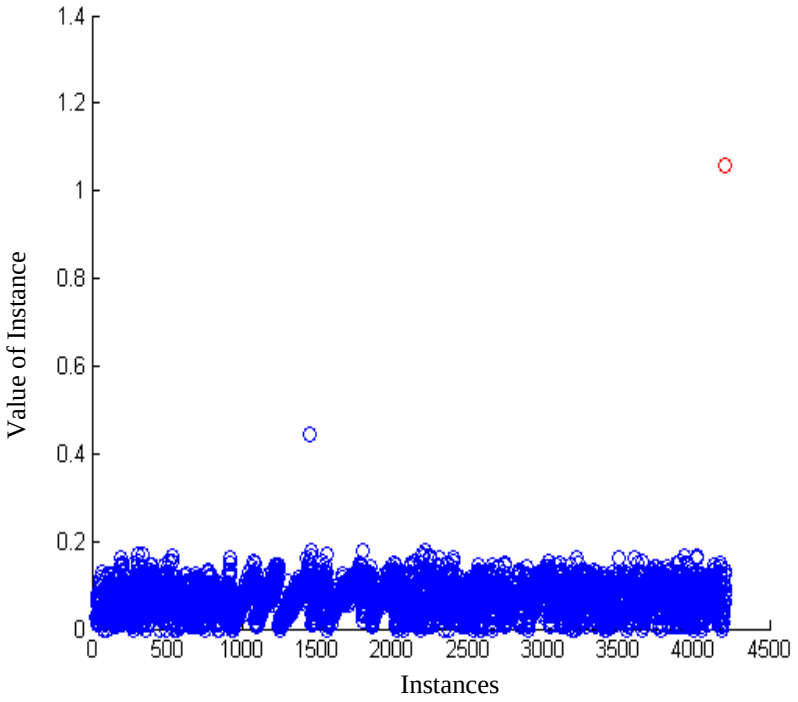


Figure 4.12: Outlier Plots for Abalone (Whole Weight)

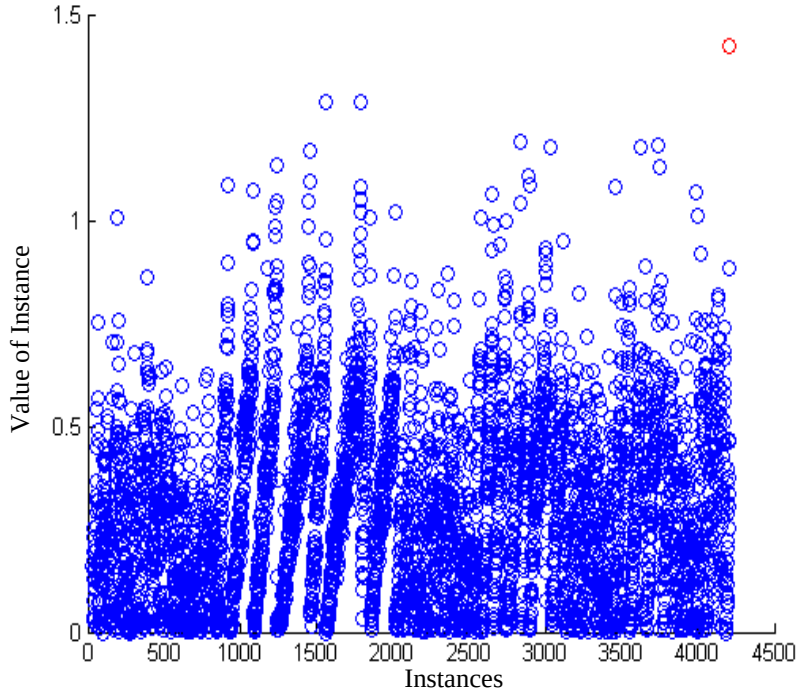


Figure 4.13: Outlier Plots for Abalone (Shucked Weight)

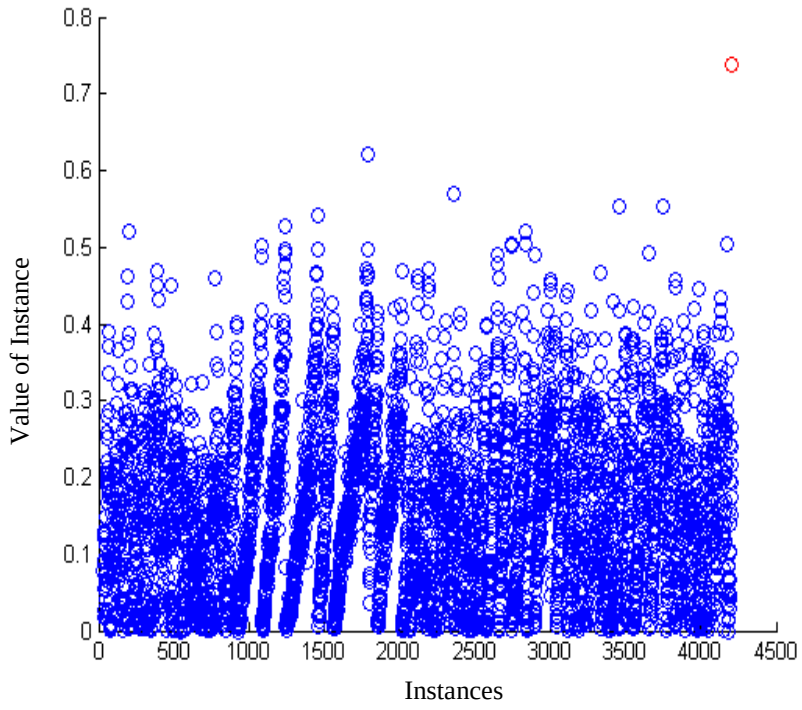


Figure 4.14: Outlier Plots for Abalone (Viscera Weight)

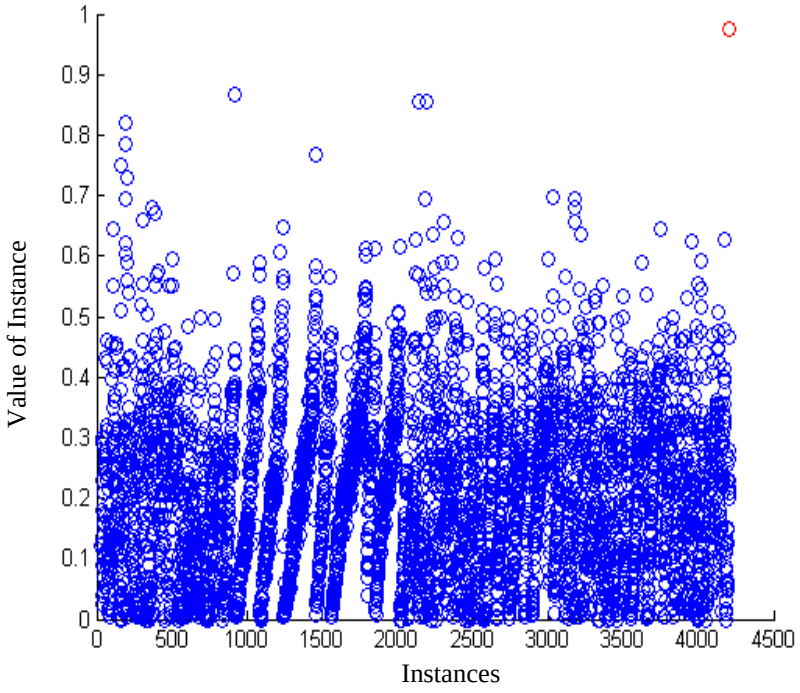


Figure 4.15: Outlier Plots for Abalone (Rings)

Based on the individual outliers for each of the separate attribute in both the datasets under consideration, a combined outlier count with varying threshold and accuracy is computed for comparison of results. The Table 4.1 presents the same.

Table 4.1: Accuracy for GWO-KNN for Iris and Abalone

Dataset	*Threshold	GWO-KNN Accuracy
Iris	0.65	96.67
	0.75	80.00
	0.85	56.67
Abalone	0.5	90.0
	0.65	53.33
	0.75	26.67

*Threshold is a numeric value for every attribute, beyond which the data point is considered as an outlier

Further, the comparison with existing approaches surrounding the Iris dataset is illustrated in Table 4.2. This table compares the results for the approaches used for outlier detection and classification. The two sets of values are representative of the error rate for the original sample without removing outliers and for the sample where the outliers have been excluded while classification. The GWO-KNN variants outperform the existing LDA, KNN and Rpart approaches as discussed in existing literature [99]. Further, amongst the GWO-KNN variants, the variant which considers 10 nearest neighbors for classification and outlier detection outperforms the others with the least error rate.

Table 4.2: Comparison of Classification Error Rate for GWO-KNN

Approach	Algorithm	Original Sample (%)	Original Sample without outliers (%)
Existing Approaches [99]	LDA	31.82	26.23
	KNN (k=7)	31.55	27.65

	Rpart	31.86	33.24
Proposed Approach	GWO-KNN (k=1)	31.01	25.48
	GWO-KNN (k=5)	30.56	28.79
	GWO-KNN (k=10)	30.48	24.67

When compared with other existing approaches for outliers in a single class for Iris (Iris Virginica), it is seen that GWO-KNN over powers the rest in terms in running time. While the precision and recall is same for flock based algorithm and GWO-KNN. Table 4.3 illustrates the results for the class Iris Virginica. Among the existing approaches including PAM, CLARA, CLARAN, Flock based and HPSO Clustering, the flock based algorithm [101] outperforms the remaining in terms of number of outliers detected which is 10. Further the precision and recall is also notable being 100% and 80% respectively. The proposed approach in this work, hybrid GWO-KNN is comparable to the best approach and takes only about 5.45 seconds to perform the outlier detection.

Table 4.3: Comparative analysis of results for class Iris Virginica

Approach	**Algorithm**	**Outliers Detected**	**Time (seconds)**	**Precision**	**Recall**
Existing	PAM [100]	7	8.56	N/A	N/A
	CLARA [100]	7	8.55	N/A	N/A
	CLARAN [100]	9	5.91	N/A	N/A
	Flock Based Algorithm [101]	10	N/A	**1.00**	**0.80**
	HPSO-Clustering [33]	N/A	6.00	1.00	0.25
Proposed	GWO-KNN	10	**5.45**	**1.00**	**0.80**

A deeper exploration justifies the reason behind the reduced time and the increased for the algorithm. The approach uses bio-inspired GWO which enables speedy convergence and avoidance of local optima thus enhancing accuracy. Further, the initial centers are obtained from the KNN approach which also directs the output to a more accurate solution.

4.1.5 Conclusion and Future Scope

The proposed hybrid GWO-KNN approach has been tested in the supervised scenario in this section of work. The datasets comprised of both regression and classification with varied number of instances and attributes. The approach seems to work well by identifying outliers with a good rate of accuracy. The threshold and the number of neighbors to consider for the underlying KNN can be varied depending on the datasets under consideration. The results have been compared for different test scenarios for various performance metrics including, accuracy, precision, recall and the number of outliers identified. The results are compared with existing heuristic and meta-heuristic variants available in the literature. Further, another test for classification after removal of the identified outliers also reveals that subsequent data mining after removing the identified outliers also improves.

In the future, the proposed approach can be further extended for application in the unsupervised domain by introducing K-Means clustering. The proposed approach can also be scaled for large datasets and different domains where there are still few evidence surrounding the application of bio-inspired algorithms. Big data frameworks including Map Reduce can be integrated with the approach for better accuracy and convergence speeds for large datasets. Domain knowledge availability and data distribution statistics can further help in improvisation of the accuracy of the proposed hybrid algorithm. Lastly, automatic identification of optimal value for 'k' and threshold can also greatly affect the final outcome.

4.2 Outlier Detection among Influencer Blogs

4.2.1 Introduction

The exponential increase in the use of internet in this era of digitization across the world has become an important source of competitive edge for the marketing of products and services [103]. This explosion of digital marketing has completely revamped the way business is done and the brand positioning strategy of the organizations [104]. Organizations have realized the importance of web visibility for better customer engagement [105]. These organizations have thus started adopting ways to artificially boost their presence on the web using digital marketing specifically opening new avenues for influencer marketing. Influencer marketing is an approach to marketing that focuses on individuals that advise the decision-making consumers. Such people are referred to as influencers and often play a critical role in the customer engagement process [106]. These influencers often need to build large amount of content in order to maximize web visibility.

The use of web analytics for enhancing digital marketing has been in practice for the last few decades. However, organizations are still not able to fully utilize the core potential of these techniques for improvising their web visibility. Studies highlight opportunities and practices in web analytics that organizations may adopt for better online marketing [107]. The optimization comprises of two primary categories of on-site (a measure of actual visitors on the website) and off-site web analytics (comprising of tools measuring website audience) [99].

One primary reason for failing to achieve the desired promotion from web analytics in online marketing is inexperienced and un-skilled influencers. These influencers in order to expedite the process use unethical practices like artificially generating keywords and links to build low quality content. This not only results in ineffective off-site analytics but also proves detrimental if detected by search engines [99] [108]. After the Google's Panda and subsequent updates, such malpractices for artificially boosting the website rank on search engines have resulted in penalization and website delisting from search engines [109]. This work thus primarily focuses on identifying outlier influencer websites for the purpose of effective off-site web analytics.

4.2.2 Background

There are several freelancing platforms including Blogmint, Influencer, Upwork and Craiglist that offer freelancers to build content on topics that may be utilized for generating back links and keywords for the customer website [110] [111]. These techniques attract traffic to the customer website and artificially boost the website rank. However, the influencers in the process to expedite the process generate low quality content that is often not original and use techniques like article spinning, keyword stuffing, link building and link farming [99] making the website quality a key driver for successful e-business [112]. The customer is often not aware of the adverse effects of such techniques and thus in the long run these may even lead to penalization by search engines.

Literature highlights several metrics that can be used to gauge the visibility of a website on search engines. Search engines use different ranking algorithms and approaches for ranking web pages in search results. The search engine data can be useful for analyzing the rank of websites [113]. Kleinberg [114] developed a link analysis algorithm for hubs and authorizes referred to as Hyperlink-Induced Topic Search (HITS) that rates web pages. Further, Page Rank was also developed, it was used by Google Search to rank websites in their search engine results [115] and Yahoo uses the Trust Flow for its link analysis to semi automate the segregation of useful and spam web pages [116]. There are no significant discussions surrounding these metrics in academic literature and majorly various link data providers like Moz, Majestic, Ahref and Webmaster tools have developed ranking mechanisms that are used worldwide for analyzing the position of a page in SERP.

Studies in literature also discuss about website selection for advertising campaigns [117]. To avoid such spam within the website, our study proposes an outlier detection approach that uses website KPIs to identify spam influencer websites that indulge in low quality content building. Metrics like page rank, page authority, domain authority, Alexa rank, Google index, social shares, trust flow, citation flow, links, external equity link; external back links, referred domains and domain age are used as indicators for identifying spam influencer websites. A spam score

is further associated with each of the 2751 websites considered for the analysis. A bio inspired wolf search and bat algorithm integrated with K-Means is used for subsequently segregating the outlier websites.

4.2.3 Methodology

This section of the work proposes a hybrid algorithm for detecting outliers in influencer blogs for the purpose of digital marketing. After the required data collection as described in Section 3.2, a statistical t-test is performed to finalize the metrics. The dataset is divided into two equal sets and 500 influencer websites each having a spam score less than 5 and greater than 5 are taken as sample for conducting a statistical t-test to identify metrics that are significantly different in the two sets. Since, the range of values of each of the metrics is considerably varied; min-max normalization is used to standardize the data to a 0-1 range. Subsequently t-test is conducted and the metrics having a p-value less than 0.05 are considered insignificant for further analysis. The final dataset for analysis thus comprises of the 12 significant attributes namely Domain Authority (DA), Page Authority (PA), Moz Rank (MR), LinksIn (LI), External Equity Links (ELL), Alexa Rank (AR), Citation Flow (CF), Trust Flow (TF), External Back Links (EBL), Referred Domains (RD), SemRush URL Links (UL) and SemRush Hostname Links (HL) for 2751 influencer websites. The p-values are illustrated in Table 4.4.

Table 4.4: Statistically significant Influencer metrics with p-value

Metric	P-value	Metric	P-value
Domain Authority	0.038	Citation Flow	7.23E-28
Page Authority	0.030	Trust Flow	0.0003
Moz Rank	1.03E-15	External Back Links	0.002
Links In	0.048	Referred Domains	3.94E-32
External Equity Links	1.25E-05	SemRush URL Links	7.12E-28
Alexa Rank	9.23E-15	SemRush Hostname Links	0.045

After the finalization of metrics as defined above outliers are detected using the proposed approach. Bio-inspired algorithms have been one of the most popular optimization techniques and mimic swarm behavior for optimization problems [14] [118]. Tang et al. [119] thus integrate a few popular bio inspired algorithms with K-means to avoid the local convergence. This study thus utilizes the integrated bio inspired wolf search algorithm for outlier detection.

The wolf search algorithm (WSA) is one such optimization approach that is said to overcome local optima by imitating the wolf preying behavior [51] [119]. In the current work, the number of clusters is identified as 2 for normal and outlier data points. The wolf population is initialized with visual distance and escape probability. The initial centroids are assigned for the two clusters. The fitness for the centroid in each wolf is calculated and the best solution is identified. The random preying behavior of the wolf is done by selecting a companion having the best solution within the visual distance. If the fitness of the companion is better than the self-fitness of the wolf the companion is selected and is thus approached. After the prey is hunted the wolf randomly selects a position beyond the visual range and the process is repeated from the new location. The centroids with the best fitness are considered as the final solution.

Further, the results are compared with the integrated bat algorithm (BA) which uses the echolocation behavior of bats to find the prey and differentiate between different insects even

in the dark [25]. The bat algorithm is one of the most popular algorithms used for several engineering, multi-objective and constrained optimization problems. For the integrated bat approach along with the two clusters, the bat population, frequency factor and loudness are initialized. The initial clusters are randomly assigned or the bat population.

For each bat, the initial centroids are similarly identified. The fitness of the centroids is computed and the best solutions are identified. Further, the new solution is generated by adjusting the frequency and velocity. If the randomly generated solution is greater than the defined pulse rate, a new best solution is selected from the best solutions from each of the bats. The new solutions are accepted by adjusting pulse rate and loudness for subsequent iterations. The pulse rate is increased and the loudness is decreased for the next iteration.

Thus the bio-inspired algorithms help in identifying the best cluster centroids over iterations. The formulation of centroids is mainly iteratively guided by the search agents in the mentioned approaches. Since the dataset considered for this study requires only two clusters and has a total of 13 attributes for which the centroid values need to be computed. The $Centroid_{ij}$ is value of the centroid for i^{th} cluster and j^{th} attribute and is expressed using Equation

$$Centroid_{ij} = \sum_{k=1}^{SolSpace} weight_{ki} datapoint_{kj} \Big/ \sum_{k=1}^{SolSpace} weight_{ki} \tag{4.1}$$

The centroids largely depend on the weight that tells whether the data point belongs to the cluster or not. $weight_{ki} = 1, if\ datapoint_k \in cluster_i; else\ weight_{ki} = 0.$

Once the best cluster centroids are identified for the two clusters of outliers and normal data points, a distance measure is subsequently used segregate the outliers.

4.2.4 Results and Analysis

The K-means integrating WSA and BA algorithms have been used in this study for detecting

outliers. The use of bio-inspired algorithms avoids locally optimum solutions. The study demonstrates the segregation of outlier influencer websites based on certain KPIs that have been extracted for a set of 2751 influencer websites using APIs. A total of 13 attributes are considered for detecting the outlier influencers for off-site web analytics. The spam score is excluded for the classification and is used for the validation.

Table 4.5 highlights the cluster centers for the remaining 12 metrics. The table lists the cluster centroids for the authentic blogs (A) and outlier blogs (O) for both WSA and BA.

Table 4.5: Cluster Centers obtained using WSA and BA

		DA	**PA**	**MR**	**LI**	**ELL**	**AR**	**CF**	**TF**	**EBL**	**RD**	**UL**	**HL**
WSA	**A**	0.12	0.19	0.18	0.69	0.32	0.74	0.55	0.76	0.84	0.58	0.46	0.75
	O	0.08	0.11	0.09	0.45	0.17	0.31	0.48	0.36	0.61	0.24	0.13	0.31
BA	**A**	0.20	0.27	0.24	0.81	0.39	0.96	0.59	0.97	0.98	0.40	0.31	0.50
	O	0.11	0.12	0.09	0.33	0.17	0.01	0.51	0.06	0.49	0.35	0.21	0.44

The results for the two approaches used for the purpose show that the bat algorithm shows higher accuracy. Out of 2751 influencer websites, 1254 websites were identified as outliers based on their spam score and manual examination. The bat algorithm correctly identified 1218 giving an accuracy of 97.12% while the wolf search algorithm correctly identified 1203 with an accuracy of 95.93%. However, time taken to converge to the optimum solution is 22.61 seconds for BA while it is just 16.18 seconds for WSA. The Figure 4.16 demonstrates the outlier plots for WSA and BA.

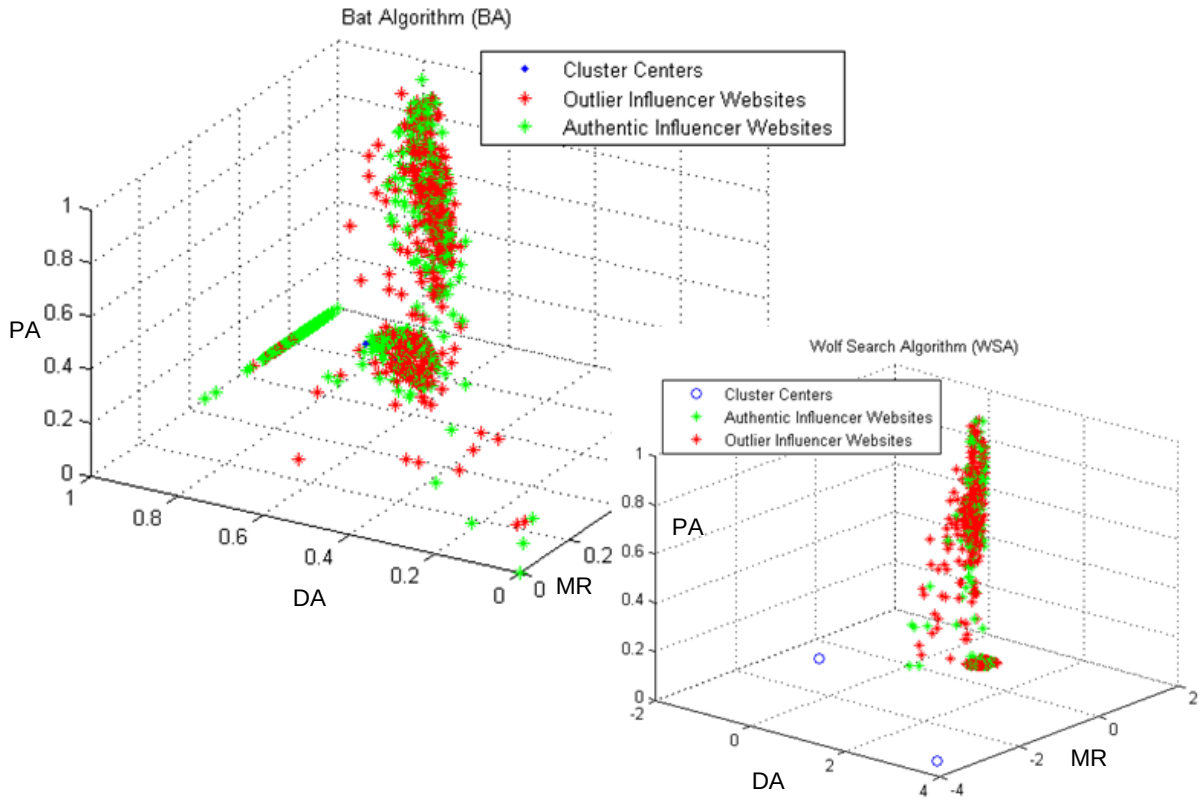


Figure 4.16: Outlier Plots for WSA and BA

Thus, the findings indicate that a large number (45.58%) of influencer websites are actually outliers. The reason behind this is that majority of influencer websites being categorized as outliers is because these blogs are heavily dependent on techniques like article spinning, link farming and keyword stuffing for content building and subsequent promotion. They often pick up original content and spin/manipulate the content by paraphrasing and including keywords related to the consumer domain to gain traction. These practices are often deemed unfit when it comes to digital marketing. However, the customers adopting these services are often not aware of such malpractices adopted by the websites. This has adverse effects on the consumer website in the long run and may even result in penalization. The use of KPIs in identifying such outlier influencers thus segregates these websites on the basis of publically available metrics from several service providers.

The results for the two approaches used for the purpose show that the bat algorithm shows

higher accuracy. Out of 2751 influencer websites, 1254 websites were identified as outliers based on their spam score. The BA correctly identified 1218 giving an accuracy of 97.12% while the WSA correctly identified 1203 with an accuracy of 95.93%. However, time taken to converge to the optimum solution is 22.61 seconds for BA while it is just 16.18 seconds for WSA.

4.2.5 Conclusions and Future Scope

With the increased internet use and online marketing opportunities, organizations have realized the importance of web visibility and have started leveraging the power of internet to reach a larger audience for their products and services. This has opened new avenues for digital marketing especially influencer marketing where on several portals have emerged to encourage these influencers to build content for customer businesses. However, this process of content building generates a lot of spam content within these websites when done in bulk for a large consumer base and often involves techniques like article spinning and keyword stuffing for user traction. Such practices are not considered ethical as per the search engine guidelines and affect the consumers adversely.

This study thus attempts to use publically available influencer website KPIs, a total of 13 attributes including Domain Authority, Page Authority, Moz Rank, Links In, External Equity Links, Spam Score, Alexa Rank, Citation Flow, Trust Flow, External Back Links, Referred Domains, SemRush URL Links and SemRush Hostname Links for 2751 influencer websites. Further, K-means integrated bio-inspired computing techniques are used for detecting and segregating outliers from the extracted data. Findings indicate that such approaches overcome local optima problems and give globally optimum solutions for such NP hard and computationally extensive data. Further, it is seen that the integrated bat algorithm gives better accuracy than wolf search algorithm as demonstrated in existing literature when the approach is used for clustering [119]. Our study re-establishes the same for the web analytics data set under consideration for outlier detection by extending the proposed approach.

This work uses KPIs and segregates outlier influencer websites that is beneficial for off-site web analytics. This may be useful for preventing consumer investments to such spam influencers that may adversely affect the websites position on search engines in the long run. Apart from the KPIs, content based analytics including keyword density, lexical diversity, meta-information and topic modeling may also be incorporated in the analysis.

Future studies can be extended to using social media analytics for further validation of the results since social media platforms are utilized by consumers for raising concerns regarding the services used by them. These platforms specially, Twitter and Facebook profiles of such influencer websites provide a lot of information in the form of user generated content that may be integrated with the existing metrics to reinforce the findings. An empirical validation of the results can also be done using a structured questionnaire for the consumers opting for such influencer marketing services and the short term and long term impact of the same on their visitors and web visibility. Existing work surrounding an analysis of results suggested by search engines for market share establishment can also be extended for influencer marketing [120].

CHAPTER 5

5. OUTLIER DETECTION IN UNSUPERVISED SCENARIO

This section of thesis is surrounding case scenarios surrounding outlier detection in an unsupervised scenario where the dataset does not have an output label for outliers. The algorithms proposed for this chapter include hybrid artificial bee colony and grey wolf optimization along with k-nearest neighbors. The section considers social media datasets from Twitter for identifying outliers in the context of buzz and fake profiles.

5.1 Identifying Buzz in Social Media

5.1.1 Introduction

Social media specifically Twitter and Facebook have become a major platform for millions of users to communicate and engage with each other. The exponential growth in the number of internet users over the last decade has greatly impacted the way marketing is done. Digital marketing has thus become a popular choice for practioners' as it not only enables instant reach and feedback but also proves to be an inexpensive way to target millions of people. These people are the primarily the ones considering the web as their primary source and means of communication and gaining knowledge [121].

Studies in existing literature thus explore the importance of these social media platforms in disseminating information [122] and further discuss the impact this information on the society as a whole [123]. As a result of the popularity and use of these social media platforms in our lives and its influence on the way business is done, existing literature explores frameworks surrounding social media marketing that catalyze the measurement and increase in return on investments for organizations using them [124] [125] [126]. This positive impact of social media has completely revolutionized business practices [127].

Further, apart from revamping business practices, these platforms when used for communication and engagement among individuals and enterprises generate large amount of user generated content usually referred to as UGC [128] which can be very well utilized for extracting gainful insights [129]. Within these gamut of conversations on the platforms and the subsequent generation of UGC there exist events that stand out in terms of the popularity that they gain among the users [130] [131]. Such discussions/events that gain higher traction than usual are often considered as “buzz”. The "buzz" discussions can be another means to explore and glean insights about these unusual events and thus enterprises have started adopting techniques that help them monitor buzz.

These techniques often enable them to attract higher number of users by fully leveraging the power and ability of social media [79] [132]. In addition to this, organizations have also become conscious of that fact that content and its popularity among the users is a key driver while gauging the business value in the social media marketing [133]. This makes the concept of viral marketing even more concrete, an electronic word of mouth (e-WOM) that grows exponentially on social media and gains huge traction [128].

Considering the trends in marketing through social media platforms the content buzz is a hype. This section of the thesis thus uses 11 metrics that comprise of created discussions, increase in authors, attention level, contribution sparseness, author interaction, author count and average length of discussions are used to model the buzz. A total 583,249 instances acquired from Twitter discussions have been used for the work. Considering the large amount of data used, future extension can be done with the use of including UGC. This calls for a need of including an approach that can handle large amounts of data, a bio-inspired algorithm is adopted to model the data attributes [14] [48]. An artificial bee colony algorithm combined with k-nearest neighbors (ABC-KNN) is thus proposed.

5.1.2 Background

Social media marketing focuses primarily on popularity of the content on the platforms and is usually referred to as content buzz. The emergence of Web 3.0 and the interactive web has enabled practitioners and marketers to communicate and engage with a larger customer base in a creative manner by adopting various strategies for internet marketing [134]. Further, these interactive social media platforms have greatly enhanced e-WOM following greater information diffusion and viral marketing [135]. Literature also investigates frameworks that have been adopted for marketing content on the web through social media [136]. The concept of content virality emerges from information diffusion which is simply an extension when the content propagates to a larger set of people at a pace faster than usual. This is then commonly referred to as "buzz" on the interactive web. Literature investigates a large number of metrics that may affect the virality of content on these platforms. This usually includes the topic being discussed, the reaction of users on it and how the network dynamics changes [137] [138].

Twitter being one of the most popular social media platforms attracts variety of users including celebrities, political figures and enterprise organizations as a platform to raise their opinion about the happenings around them [139]. Thus, content buzz on platforms like these spreads like a ripple. This buzz may be anything with higher than usual user attention including popular events and people's opinion about the same [140] [141]. Studies also investigate the drivers that may be contributing to this social media content buzz and the asset value that comes with it [142] [143].

Further, coming to the domain contribution of the work, it is evident that outlier detection approaches have been popular in domains of cyber security with applications surrounding intrusion detection, fraud detection and wireless sensor networks [4]. However, the literature lacks evidences of adopting outlier detection approaches in the domain of information propagation and diffusion and how they may be clearly defined. With the growing trend about content popularity and buzz, this work is an attempt to explore content buzz by using eleven attributes collected for over 5 million discussion instances on Twitter. The methodology

comprises of adoption of bio-inspired computing where in the behavior of biological species is used for arriving to workable solutions in complex multi-dimensional problems.

5.1.3 Methodology

The work adopts a mixed methodology with a combination of social media analytics and bio inspired algorithms as it seemed challenging to address the objective otherwise. The work primarily focuses to identify content buzz in Twitter discussions using a hybrid bio-inspired computing approach proposed for detecting outliers. The "buzz" discussions in this context are considered as potential outliers throughout the analysis. Section 3.3 provides insights about the nature of data under consideration.

The distribution of data for the dataset is illustrated in Figure 5.1 with discussions identified as buzz highlighted in red. The break point evidently differentiates the buzz and non-buzz discussions with content buzz being captured with values greater than $\mu + 2\sigma$. The discussions are plotted by arranging the buzz in increasing order and are representative of the content buzz beyond the desired threshold. Therefore, any Twitter instance with number of active discussions beyond $\mu + 2\sigma$ is demarcated as "buzz" in the current work. The remaining instances are the "non-buzz" instances when seen through the lens of activity and visibility using the identified model.

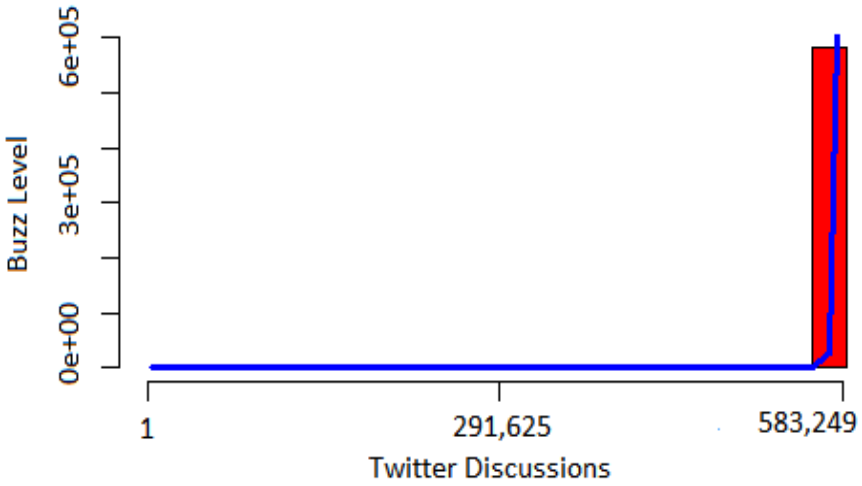


Figure 5.1: Graphical distribution of buzz

(Based on activity incited by individual tweets)

This section of the thesis models the final dataset with 11 attributes for mining outliers in the form of buzz discussions. These are subsequently validated using the output variable for computed the accuracy of the proposed hybrid bio-inspired approach. A min-max normalization is further employed before the outlier detection since the values for each of the 11 attributes lie in varied ranges. The normalized data lies in the range of 0 and 1 for every attribute being modeled. Post the normalization of the data, it is randomly sampled for training and testing and then modeled using the proposed hybrid algorithm for identifying buzz discussions. The following sub-section illustrates the proposed bio-inspired approach used for mining outliers.

Hybrid Artificial Bee Colony (ABC) Approach

The current work combines artificial bee colony optimization (ABC) with k-nearest neighbors to identify outliers in the form of buzz. The ABC optimization uses a population based search mechanism where artificial bees look for food sources [23]. The ABC algorithm has been modified over the years to be used for different application domains [145]. The algorithm is known to produce promising results while optimizing an objective function [146]. The algorithm is inspired from the intelligent behavior of honey bees. The bee colony comprises primarily of three different categories of bees depending on how they search for food sources, these include the onlookers, scouts and lastly the employed bees.

Further, ABC is popular yet simple optimization algorithm that results in a globally optimum solution to the problem under consideration by varying colony size, the number of food sources, and the foraging cycles. These acts as control variables and help in reaching to the optimal solution by varying values based on the objective function. The ultimate goal of the honey bees is to search for food sources having the highest amount of nectar. Equation 4.1 represents the candidate solution (CFp_{ij}) and is used for the movement of bees in the direction of the food source. The updated position in the multi-dimensional search space is represented as:

$$CFp_{ij} = pos_{ij} + \tau_{ij}(pos_{ij} - pos_{kj}) \tag{5.1}$$

where $k \in [1,2 \ldots . \in n]$,n = employed bees, $j \in [1,2 \ldots . P]$, P is the optimization parameters, reflecting the dimension of the solution and τ_{ij} is a number randomly generated between [-1,1], controlling the neighborhood of pos_{ij}. The difference $(pos_{ij} - pos_{kj})$ decreases every subsequent iteration towards the optimum solution.

The way the food source is identified is dependent on the type of bee that is in search of it. The employed bees use their own experience for locating the food sources, while the onlookers use the experience of employed bees who dance to reflect the position of the food source to the onlookers. The scouts however do not use any kind of experience for selection of food sources and locate the same. Once the new food source position is identified, it is memorized only in the scenario when the amount of nectar at the position is higher than the previous best position (having the highest amount of nectar).

The solution quality often indicative of the fitness of the solution refers to the nectar amount ($NecAmt$) present at the food source. The ecosystem consists of one employed bee for every food source present in the bee hive. The employed bees modify the current position in the memory based on the locally available visual information about the nectar amount. The same is done by testing the amount that reflects the fitness of the food source being considered. The

position is updated if the nectar amount is greater than that at the previous position. The food source selected by the onlooker bee depends on the probability that corresponds to the food source Fp_i , computed using:

$$Fp_i = \frac{NecAmt_i}{\sum_{n=1}^{FdSrc} NecAmt_i} \tag{5.2}$$

$NecAmt_i$ represents the solution's fitness which is indicates the amount of the nectar available at the position i. The $FdSrc$ is representative of the onlookers available. Further, the fitness varies with the problem in consideration and refers to the error emerging from the data points clustered together (being a minimization problem in this case). Figure 5.2 illustrates the hybrid approach proposed in this work, KNN integrated artificial bee colony approach (ABC-KNN).

Pseudo-code of k-nearest neighbor integrated artificial bee colony approach

Begin

Initialize the population, colony size, number of food sources and foraging cycles.

Initialize the initial cluster centers randomly

Repeat

For each set of data $k = 1 \ldots s$

/ 's' sets of employed bees */*

do */*onlooker bees */*

*/*Finding the best food source using the experience of employed bees*/*

Pass the new food source position (CFp_{ij}) to k-nearest neighbor fitness function to minimize F_n=$d(Ti, p) = \sqrt{\sum_{i=1}^{n}(Ti - p)^2}$, where Ti is the cluster centroid of i^{th} cluster and p is the data point under consideration (the new food source) and get the output.

Assign the (data point) discussion to the cluster (normal or buzz) based on the distance

βi is the best instance, data point having minimum distance *(dist)* to the assigned cluster $[min(\sum_{k=1}^{s} dist_k)]$ for buzz. The βi has the dimensionality equal to number of attributes.

The βi is thus updated iteratively.

Move to the next set of data instances

End For

For each instance

Mark data point ($datap$) as outlier if it lies beyond outlier threshold computed using βi and standard deviation.

End

Figure 5.2: Pseudo-code of the proposed ABC-KNN approach

The proposed model is compared GWO-KNN which is based on the hunting behavior of wolves [26]. In addition to this, the results are also compared with a traditional machine learning approach used on the same dataset. The study uses a Regression Random Forests to identify buzz.

5.1.4 Results and Findings

The proposed integrated ABC-KNN as illustrated is used to segregate the outliers using mentioned set of attributes. The result validation is done through the output variable which is indicative of buzz, the mean number of active discussions. This variable is predicted using the 11 attributes when modeled together through the hybrid approach. The Twitter instances with value of output variable greater than the defined threshold of $\mu + 2\sigma$, i.e. the ones lying beyond the significant 95% of the dataset are considered as outliers. The proposed hybrid ABC-KNN approach results into an accuracy of 98.37%. The plots for outlier buzz instances are illustrated in Figure 5.3. The data points marked in red are indicative of the "buzz" discussions while the ones highlighted in blue represent the non-buzz discussions that did not gain higher than usual user attention through the UGC.

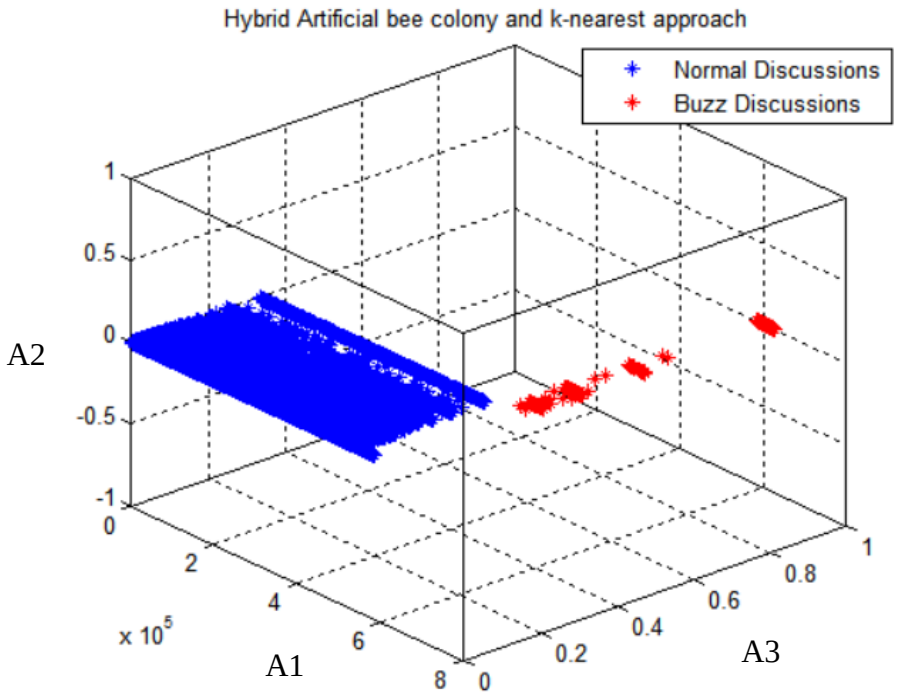


Figure 5.3: Content Buzz Identification using ABC-KNN

Further, the results are also validated using a cross validation mechanism with a five-fold validation giving an overall accuracy of 97.87%. A 60:40 ratio is taken for training and testing

respectively. The random sample is for training is testing is taken five times for the cross-validation process to ensure the sanctity of results. Table 5.1 is representative of the outlier thresholds for the entire attribute over the five validation steps that have been computed through the proposed approach for outlier detection. The obtained vectors are utilized to segregate outliers. The validation is also done by comparing the buzz discussions achieved as output by modeling the 11 attributes and the points beyond the threshold for the output variable representing mean number of active discussions.

Table 5.1: Outlier Threshold for Buzz using 5-fold Cross Validation

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11
I	0.82	0.52	0.60	0.81	0.75	0.36	0.81	0.42	0.78	0.46	0.47
II	0.79	0.52	0.67	0.84	0.72	0.40	0.72	0.42	0.77	0.46	0.40
III	0.80	0.50	0.78	0.88	0.73	0.48	0.73	0.62	0.76	0.61	0.39
IV	0.72	0.46	0.68	0.93	0.66	0.41	0.86	0.35	0.69	0.37	0.46
V	0.92	0.53	0.57	0.90	0.69	0.36	0.88	0.37	0.73	0.39	0.46

When compared to the similar grey wolf optimizer variant, the proposed approach outperforms the same in terms of accuracy (98.37%) which is 97.12% for GWO-KNN. Further, on comparing the convergence speeds of the two approaches, it is noted that the proposed ABC-KNN converges to an optimum faster as depicted in Figure 5.4.

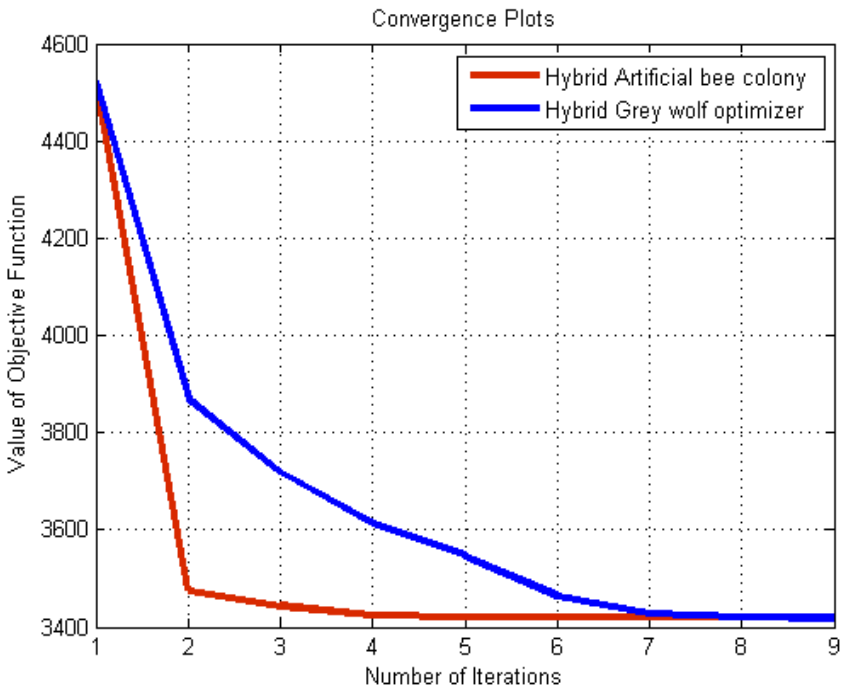


Figure 5.4: Convergence plots of the hybrid ABC-KNN and GWO-KNN

The proposed ABC-KNN approach for outlier detection thus outperforms the GWO variant in terms of both the accuracy with which the outliers are identified and the convergence speed for reaching to a global optima. A very famous optimization theorem, the No Free Lunch (NFL) theorem provides us with an interesting finding which states that there can be no optimization algorithm that can be considered best [147]. An algorithm may perform exceptionally one type of dataset and on the other hand, may fail miserably for the other. The main reason for this is considered to be the data distribution and the type of data being handled.

In the current scenario, the proposed approach is however successfully able to identify buzz discussions and attaining a globally optimum solution. With the huge amount of data pouring in the form of UGC and other sources, such optimization techniques shall be useful in reaching to solution faster. This overcomes the drawback of the traditional machine learning approaches that tend to lack accuracy and speed in such scenarios because of the large amount of data and computational complexity required to analyze the same. The comparative results with existing traditional regression random forests approach is depicted in Table 5.2.

Table 5.2: Comparison of ABC-KNN and Regression Random Forests for Twitter Buzz

Approach	Algorithm	Accuracy
Proposed	ABC-KNN	97.8%
Existing [144]	Regression Random Forests	94.2%

5.1.5 Conclusion

The advent of Web 3.0 has led to an increased usage of social media platforms and interactive media. The utilization is not only restricted to communication and interaction among individuals. These platforms are also being used for purpose of digital marketing and consumer engagement. The underlying concept of content buzz has thus been trending and any piece of content gaining popularity higher than the usual is of particular interest. Existing studies in literature have captured various attributes surrounding content popularity and virality.

This work on the hand attempted to identify buzz in social media specifically Twitter using a set of 11 attributes comprising of increase in authors, attention level, created discussions, burstiness level, author interaction, author count, contribution sparseness and average length of discussions are modeled to identify buzz discussions and segregating them from the remaining discussions that have not gained higher user attention. A total of 583,249 Twitter discussion instances have been utilized for the analysis.

Talking about the methodological contribution of the work we propose a hybrid ABC-kNN approach for identification of outliers as buzz discussions. The approach integrates the traditional KNN with ABC. The performance of the same is better than traditional heuristic approaches as it converges to a globally optimum solution avoiding the chances of getting stuck into a locally optimum solution that the traditional approaches are often prone to. The model considers buzz discussions as outliers deviating from the normal and detects them successfully resulting in an accuracy of 98.37%. The approach has been compared with similar hybrid GWO-KNN for detecting outliers and outperforms the same in terms of accurate identification

of outliers and convergence speed. The proposed hybrid approach can also be adopted for similar domain specific applications including email and social media spam identification, anomalies in purchase behavior, websites and similar domains that demand workable solutions with a constraint on time and complexity.

Findings of the work may have practical implications in various domains including e-commerce, digital marketing and e-governance to model which metrics may be responsible for creating buzz and their subsequent impact. Future scope of the work can extend the set of metrics including content, descriptive and network analytics attributes to model popularity over platforms. This can include analyzing UGC for sentiment and network parameters. The approach can also be scaled for datasets having big data properties (greater volume, variety and veracity) through frameworks for parallel programming.

5.2 Detecting Fake Profiles on Social Media

5.2.1 Introduction

Over the past decade social media has gained immense popularity with Facebook, Twitter, LinkedIn and Instagram being the most widely used platforms [64]. With the information availability on these social platforms, users are compelled to seek and strive for traction by depending on other users in the network to propagate their content [92]. Both individual users and enterprises are thus trying their best to take advantage of these interactive platforms for expansion of their content scope. This is usually done by trying to propagate the content resulting into virality over the social network. Therefore, many organizations are adopting various digital marketing strategies where they aim on increasing set of users that will catalyze their content propagativity.

This results in adoption of unethical approaches, which results in creation of plethora of fake user profiles that are solely created for the purpose of artificially boosting the follower count on these social networks. This in turn largely affects the favorable social votes, likes and shares for the individual or the organization in quest of the same. Studies report a large percentage of Facebook (44%) and Twitter (33%) followers to be fake. In fact, New York Times reports evidences of fake followers becoming business on Twitter worth millions of dollars. Twitter users both individuals seeking attention or organizations aiming for greater outreach are paying hefty amounts for large follower lists. Literature showcases evidences of studies that take into consideration followers for analysis of data which makes it important to segregate fake artificial followers from the authentic ones.

There are existing studies that have attempted to identify fake Twitter profiles emphasizing the fact that opinions shared by these fake followers can be unreliable and misleading for the community [149] [150]. These studies use a limited set of metrics and show no correlation with the personality dimensions of the users considered. Such an interdisciplinary approach for detecting fake Twitter profiles is still unexplored. Further, the use of bio-inspired algorithms

enhancing the accuracy and convergence speeds for reaching to a solution are an added methodological advancement along with the domain improvisation.

5.2.2 Background

With wide use of internet by the masses, the 21st century is witnessing an explosion of user generated content on the web [74]. This user generated content available on the social media networking platforms can be used in different fields like marketing, e-commerce, finances and so on by gauging the user's action and response to the events that occur. The information acceptability and content availability become major factors in influencing user behavior due to which social media has thus become a source of communication and engagement with stake-holders [75] [76] [77].

In recent times, organizations are also deploying resources to manage social media as it constitutes a substantial part for improving organic search results [70] [71]. This helps to direct potential customers to websites from search results and also from high integration with social media users [72]. The world of Web 3.0 and the huge influx of information make it infeasible to focus on all channels since business needs and channel receptivity depends on it. Considering this large impact and utilization of social media in varied domains it becomes critical to validate the authenticity of this shared content. Thus identifying outliers in the domain becomes essential.

Existing studies discuss applications surrounding community detection in this domain and outliers could be mined [37]. These outlier profiles are often used by marketers to boost and promote content on social media. Organizations are paying hefty amounts to marketing agencies fake followers, artificially paid likes, comments and shares to gain user traction. It therefore becomes important to identify such fake profiles as there is a high probability that the information shared and propagated by these may be misleading or unauthentic. There are extensive studies that identify fake profiles in these social networks, specifically focus on detection of spam specifically on Twitter [148] [149] [150]. However, none of the existing

studies focus metrics extracted from UGC and do not take personality into consideration. This section of the thesis thus attempts to identify fake Twitter followers based on certain descriptive metrics mapped to their personality dimensions.

5.2.3 Methodology

A mixed research methodology had to be followed in this study essentially because a single interdisciplinary approach appeared difficult to address the research objectives. Several methods were adopted from domains as diverse as social science, social media analytics and bio-inspired computing [14]. The analysis was heavily dependent on using tweets for the analysis. The details of the data collected are expressed in Section 3.4.

After conducting the Delphi study the final metrics relevant for the analysis are identified. The approach is one of the popular alternatives to for achieving consensus to a problem [151]. The approach uses a voting mechanism within selected experts in the desired domain. The applicability of final set of parameters that may be used for the identification of nature of profiles is done by adopting this approach in the current scenario. The consensus was obtained in a total of 2 iterations with 5 experts participating in the study. The experts had diverse backgrounds including computer science, social statistics and psychology and have years of working experience in behavior analysis in social media. The Delphi consensus resulted in the finalization of 12 relevant parameters categorized into five personality dimension as depicted in Figure 5.5.

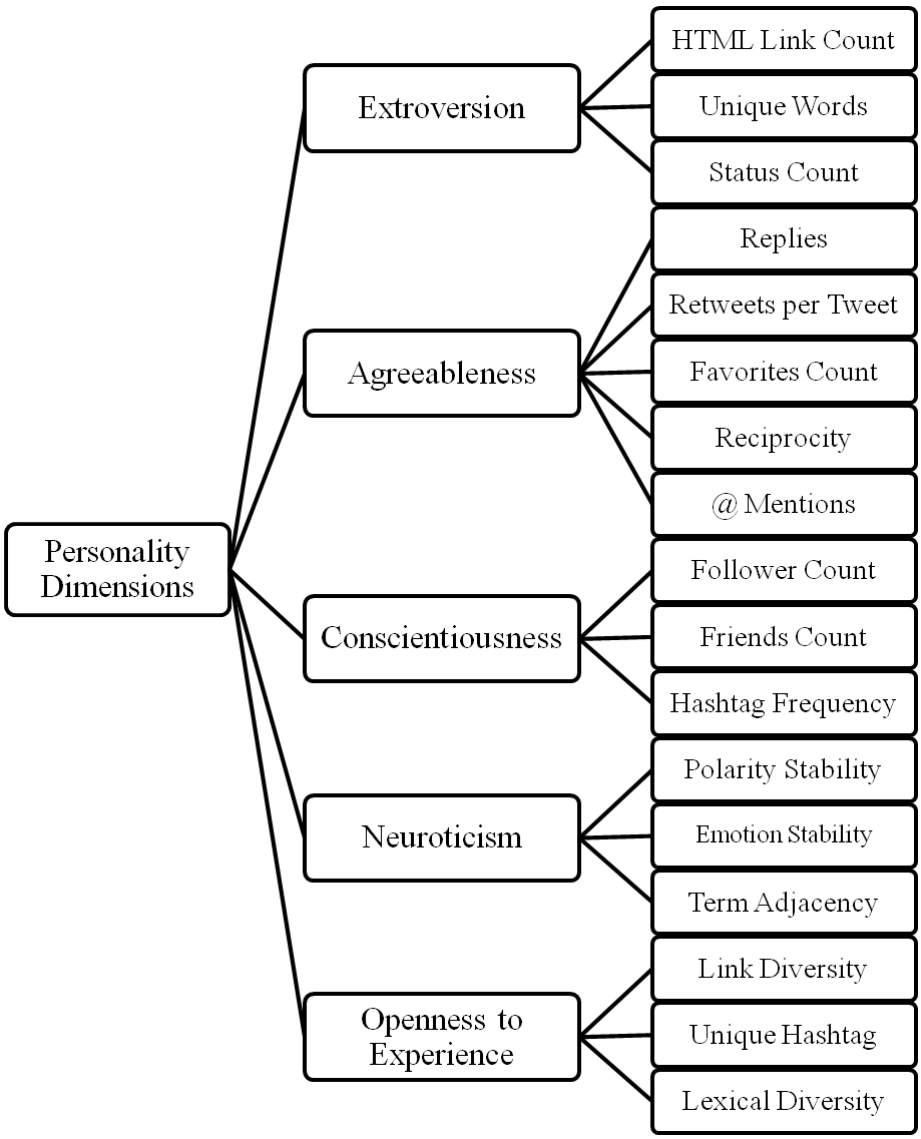


Figure 5.5: Twitter Metrics and Personality Dimension Categorization

The final set of metrics comprised of hashtag frequency, HTML link count, unique words, follower count, statuses count, @ mentions, friends count, favorites count, polarity stability, unique hashtags, emotion stability and lexical diversity. These metrics have also been grouped into personality dimensions as illustrated above. A detailed description of each of these metrics is shown in Table 5.3.

Table 5.3: Description of the Twitter Metrics for Fake Profile Identification

S. No.	Metric Name	Metric Description	Personality Dimension
1.	Status Count (M9)	A count of statuses updated by the user.	Extroversion
	Unique Words (M6)	A metric that measures average number of unique words in each status.	
	HTML Link Count (M5)	A metric that measures average number of HTML links in each status.	
2.	Favorites Count (M10)	A count of number of tweets a user has liked.	Agreeableness
	@ Mentions (M7)	A metric that counts @ mentions of other users in 50 recent tweets.	
3.	Friends Count (M11)	A count of number of users the user follows.	Conscientiousness
	Hashtag Frequency (M3)	A metric that measures average number of hashtags in the last 50 tweets.	
	Follower Count (M12)	A count of number of followers of the user.	
4.	Emotion Stability (M1)	A metric that measures the deviation in emotional score for the given emotions-joy, sadness, surprise, disgust, fear and anger for a user in the last 50 tweets.	Neuroticism
	Polarity Stability (M2)	A metric that measures the deviation in polarity score for both negative and positive polarity for a user in the last 50 tweets.	
5.	Unique Hashtags (M4)	A metric that count the number of uniquely used hashtags in 50 recent tweets.	Openness to Experience
	Lexical Diversity (M8)	A metric that defines the fraction of unique words upon the total words used by a user in 50 recent tweets.	

The Twitter data collected for the purpose of this work is highly enriched and contained, tweets, user information, html links, followers, status count, demographic location, @mentions and hash tags. The 10,000 users for which the data was collected had an average of 21,895 tweets

in their life time (a part of which was analyzed), an approximate 3285 followers per user and followed around 904 users. Some were old Twitter users, tweeting since 2007, while some were recent users that created their account in June 2016 when the data for this study was collected. The users belonged to diverse demographic locations. This section focuses on these descriptive statistics as they are widely used for analyzing the personality traits of the users. The tweet metrics like word, follower, following and status count that are finally used for the purpose of extracting intelligence give us a holistic yet simple picture of the data.

These statistics provide an insight surrounding the personality of the users and are extensively used in literature for extracting user information [153]. The tweets analyzed for this work are enriched with a number of smiley, URLs, images, mentions to other and diverse list of hashtags. The final dataset is expressed by a total of 12 attributes obtained from the Twitter profiles of the users under consideration. These computed metrics are further grouped into personality dimensions adopted from the big five framework using 555,684 tweets [152].

Methodologically, this work uses integrated approaches for detecting the outliers. This work uses grey wolf optimization [26] and artificial bee colony [23] [140] approaches integrated with KNN. The ABC produces promising results for solving single objective numerical problems. The algorithm works well for both minimization and maximization problems [47]. The GWO on the other hand, is popular for its high exploration and exploitation function that assists it in outperforming other trainers for classification scenarios [26]. Both these approaches have been effective in numeric optimization problems.

Grey Wolf Optimization

The GWO is inspired from the hunting of wolves, it identifies the most optimal solution (α), second best (β) and third best solution (δ) based on the wolves [26]. The remaining solutions, omega (ω) reflect the remaining wolves in the pack. The way wolves hunt the prey (or the optimization reaches a workable solution) is often governed by the leader, the α wolves and the remaining pack of wolves follow the leader. The optimization approach uses the encircling of

prey and subsequently hunts it for reaching to a global optima over iterative steps. The same can be mathematically modeled as:

$$\vec{X} = \left|\overrightarrow{V}.\ \vec{P}_{prey}(t) - \overrightarrow{P}(t)\right| \quad (5.3)$$

$$\vec{P}\,(t+1) = \vec{P}_{prey}(t) - \vec{U}\,.\vec{X} \quad (5.4)$$

The $\vec{X}$ is used for computation of the distance between the prey location $\vec{P}_{prey}(t)$ in that iteration (t) and the wolf location $(\overrightarrow{P}(t))$. The wolf position/location is updated iteratively based on the vector $\vec{X}$ (calculated using Equation 5.3) in the next iteration$(t+1)$. $\vec{U}$ and $\overrightarrow{V}$ are coefficients and are calculated using Equation 5.5 and 5.6.

$$\vec{U} = 2\,\overrightarrow{u}\,.\vec{d}_1 - \vec{u} \quad (5.5)$$

$$\vec{V} = 2.\vec{d}_2 \quad (5.6)$$

The coefficient vectors are computed using vector $\vec{u}$ which decreases from 2 to 0 in a linear fashion and the two random vectors $\vec{d}_1$ and $\vec{d}_2$ which vary in the range [0, 1].

On encircling the prey, the distance vector is calculated for each of the wolf using α, β and δ wolves as the location of the prey. The equations (5.7), (5.8) and (5.9) are descriptive of the step size while the equations (5.10), (5.11) and (5.12) are used for computing the final positions of the remaining wolves.

$$\vec{X}_\alpha = \left|\vec{U}_1.\ \vec{P}_\alpha - \vec{P}\right| \quad (5.7)$$

$$\vec{X}_\beta = \left|\vec{U}_2.\ \vec{P}_\beta - \vec{P}\right| \quad (5.8)$$

$$\vec{X}_\delta = \left|\vec{U}_3.\ \vec{P}_\delta - \vec{P}\right| \quad (5.9)$$

$$\vec{P}_1 = \vec{P}_\alpha - \vec{V}_{1.}\,(\vec{X}_\alpha) \quad (5.10)$$

$$\vec{P}_2 = \vec{P}_\beta - \vec{V}_{2.}\,(\vec{X}_\beta) \quad (5.11)$$

$$\vec{P}_3 = \vec{P}_\delta - \vec{V}_{3.}\,(\vec{X}_\delta) \quad (5.12)$$

where, every dot product

$$\vec{A}.\vec{B} = a_1 b_1 + a_2 b_2 \ldots\ldots + a_n b_n \quad (5.13)$$

It is assumed that the best, second best and third best wolves have knowledge about the position of the prey. Therefore, these groups of wolves help the remaining ω wolves for updating the positions based on the best search wolves in the subsequent iteration$(t+1)$.

$$\vec{P}\,(t+1) = \frac{\vec{P}_1 + \vec{P}_2 + \vec{P}_3}{3} \quad (5.14)$$

The work uses GWO-KNN to segregate fake Twitter profiles. The proposed approach for outlier detection in the form of fake profiles is utilized in the work. The results are compared with ABC-KNN. Figure 5.6 illustrates the pseudo-code for the proposed approaches.

GWO-kNN	ABC-kNN
Begin *Step 1: Initialization. Start iteration counter with t=1* *Initialize population of grey wolves randomly;* *Initialize vectors* $\vec{u}, \vec{U}$ *and* $\vec{V}$ *Initialize search agents* $\vec{P}_\alpha$, $\vec{P}_\beta$ *and* $\vec{P}_\delta$ *Determine number of groups, K.* *Assign K clusters for each wolf.* *Assign K objects as initial centroids for each group.* *Step 2: Exploration. Compute fitness of centroids for each group* *While t < MaxGeneration do* *Sort the population according to their fitness.* *Update position of each search agent* *Update vectors* $\vec{u}, \vec{U}$ *and* $\vec{V}$ *Evaluate fitness of each search agent* *Step 3: Centroid Update. Sort the population and find current best solution.* *t=t+1.* *Reassign the clusters.* *Output best cluster configuration based on best fitness.* *End while* *End*	*Begin* *Step 1: Initialization. Start iteration counter with t=1* *Initialize population of bees randomly;* *Initialize colony size and food sources.* *Determine number of groups, K.* *Assign K clusters for each wolf.* *Assign K objects as initial centroids for each group.* *Step 2: Exploration. Compute fitness of centroids for each group* *While t < MaxGeneration do* *Sort the population according to their fitness.* *Produce updated solutions* (FdS_{ij}) *for the employed bees* *Calculate probability* $Prob_i$ *for onlookers* *Evaluate fitness of the obtained solutions* *Identify abandoned scout solutions for replacement* *Step 3: Centroid Update. Sort the population and find current best solution.* *t=t+1.* *Reassign the clusters.* *Output best cluster configuration based on best fitness.* *End while* *End*

Figure 5.6: Pseudo-code for Fake Profile Identification using GWO-KNN and ABC-KNN

5.2.4 Analysis and Findings

This work address the problem of fake profiles that are increasing exponentially on social media platforms specifically Twitter. The proposed approach along with identification of these outlier profiles also maps them to logical categorization of personality dimensions. This final dataset comprises of 10,000 profiles modeled using 12 metrics using a total 555,684 tweets that are used for computation of these metrics. For the purpose of standardization and avoiding metric bias, a min-max normalization approach is used to normalize the varying values into comparable range [0, 1], for all the metrics.

Further, this normalized dataset is subsequently used to detect fake Twitter profiles by using the proposed outlier detection techniques. For the purpose of validation of the model and the applied algorithm, the dataset is initially divided into training and testing. The training data is labeled into “fake” and “authentic” profiles based on the sources of tweets liked, re-tweeted and embedded links. With a deeper look into the “fake” profiles, it was evident that these profiles repetitively promoted content from similar sources. On the other hand, the “authentic” ones showcased variety in the content re-tweeted and posted.

On examining these outlier profiles identified as "fake" using the approach, it was observed that most of these users did not have a profile picture (58%); some had celebrity pictures as their profile pictures (24%) and these celebrities already possessed verified accounts. These users also had a disproportionate follower to following ratio and did not reveal any relevant information about the user. In addition to this, these profiles barely had any original tweets and sustained from retweets to similar profiles sharing the same content repeatedly. The profiles also shared plethora of marketing web links from similar websites (79%).

The outlier plots obtained for the two approaches used for identification of such profiles have been illustrated in Figure 5.7. Findings indicate that GWO-KNN results in 1311 outlier profiles while ABC-KNN identifies 1368 profiles as outliers. On manually checking each profile, a total of 1409 profiles appeared to be “fake”.

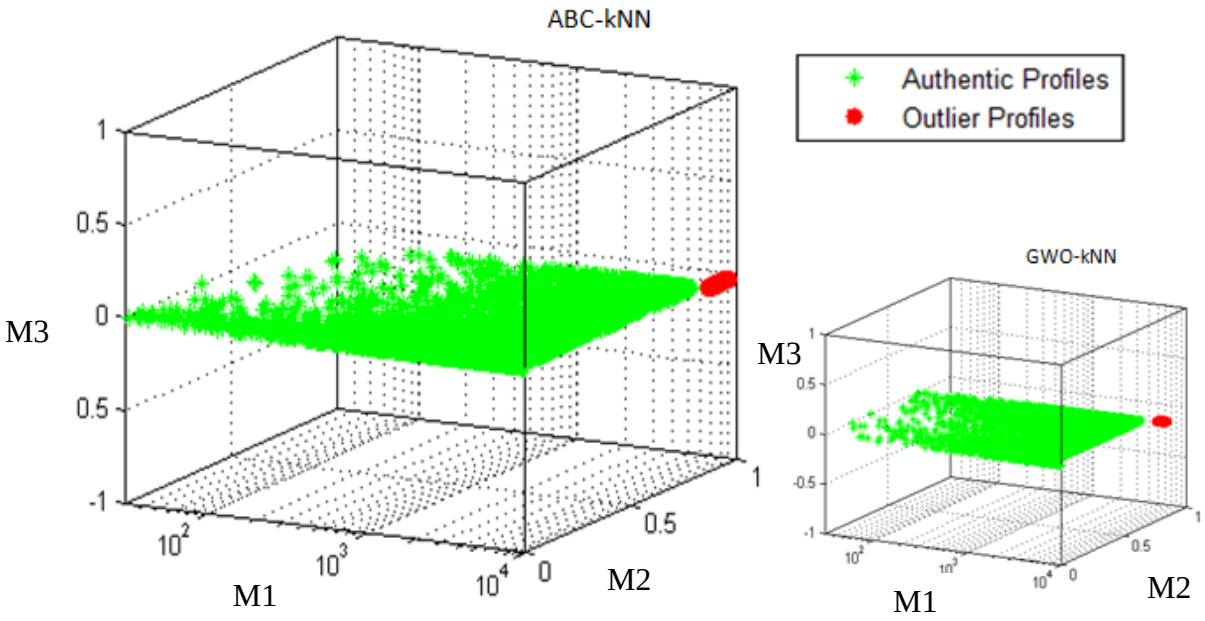


Figure 5.7: Outlier Fake Profile Plots using ABC-KNN and GWO-KNN

The results of ABC-KNN depicted higher accuracy (87.09%) when compared to GWO-KNN having 83.04%. The ABC-KNN outperforms the GWO-KNN approach for the given dataset extracted from Twitter. The GWO-KNN however reaches an optimal solution faster than the hybrid ABC approach. The same has also been identified in literature as a major pitfall of the GWO with the positions of search agents depending on α, β and δ. These computed three search agents are prone to increase the pressure of convergence to a globally optimum solution. This forces the algorithm to converge prematurely which loses the diversity of the algorithm with a compromise on the accuracy [154].

The work addresses the identification of metrics that influence in detection of fake profiles. The 12 factors that came out to be relevant in identifying the fake profiles are status count, unique words, html link count, favorites count, @ mentions, friends count, hashtag frequency, follower count, emotion stability, polarity stability, unique hashtags and lexical diversity. These are further divided into five personality dimensions to which they contribute.

Table 5.4 highlights the results for the two approaches GWO-KNN and ABC-KNN for cross validation an average (Avg.) over 10 iterations taking a set of 5000 different users in each one. The table also depicts the cluster centers using the entire set of 10,000 users (10k). For both the set of values, the cluster centers for both the groups Authentic Twitter Users (A) and Fake Twitter Users (F) have been highlighted.

Table 5.4: Cluster Centers for Authentic and Fake Profiles

Algorithm			M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
GWO-kNN	Avg.	A	0.207	0.066	0.090	0.083	0.163	0.245	0.332	0.339	0.241	0.134	0.374	0.290
		F	0.756	0.884	0.850	0.853	0.766	0.695	0.828	0.723	0.558	0.640	0.726	0.724
	10k	A	0.384	0.050	0.093	0.072	0.129	0.199	0.332	0.323	0.326	0.107	0.535	0.500
		F	0.485	0.906	0.857	0.872	0.765	0.667	0.750	0.611	0.616	0.602	0.881	0.914
ABC-kNN	Avg.	A	0.061	0.054	0.056	0.082	0.147	0.150	0.259	0.136	0.151	0.107	0.171	0.085
		F	0.881	0.846	0.903	0.855	0.731	0.579	0.831	0.701	0.607	0.543	0.627	0.482
	10k	A	0.055	0.067	0.043	0.069	0.134	0.125	0.196	0.100	0.160	0.094	0.211	0.081
		F	0.875	0.931	0.954	0.808	0.682	0.536	0.873	0.691	0.673	0.512	0.695	0.464

For the purpose of validation of the proposed methodological model for identifying fake profiles, an independent samples test is conducted. This test is used for the significance of the 12 metrics over 10 cross validation iterations for the two proposed approaches. The two population groups of Twitter profiles comprising of 20 cluster centroids in each, with a total of 40 cluster centroids combined are used for the test. The work also uses two statistical tests, one for the equality of means and the other one for the equality of variances for validation of results.

The two-tailed t-test for equal means, tests the null hypothesis such that the means of two populations are equal [155]. Such tests are usually referred to as Student's t-tests. For, the Levene's test the null hypothesis is that the population groups have equal variances [156]. The significance level for both the tests is taken to be 0.05 and the metrics that are greater than the significance level are considered to be not significant. The
Table 5.5 highlights the results of both the independent sample tests.

Table 5.5: Independent Samples Test Analysis Results

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
M1	Equal variances assumed	1.615	.211	13.873	38	.000	.68465	.04935	.58474	.78456
	Equal variances not assumed			13.873	34.900	.000	.68465	.04935	.58445	.78485
M2	Equal variances assumed	7.553	.009	21.553	38	.000	.80525	.03736	.72962	.88088
	Equal variances not assumed			21.553	19.639	.000	.80525	.03736	.72723	.88327
M3	Equal variances assumed	6.858	.013	23.874	38	.000	.80395	.03367	.73578	.87212
	Equal variances not assumed			23.874	24.360	.000	.80395	.03367	.73450	.87340
M4	Equal variances assumed	9.675	.004	50.837	38	.000	.77105	.01517	.74035	.80175
	Equal variances not assumed			50.837	21.670	.000	.77105	.01517	.73957	.80253
M5	Equal variances assumed	13.373	.001	28.235	38	.000	.59370	.02103	.55113	.63627
	Equal variances not assumed			28.235	24.621	.000	.59370	.02103	.55036	.63704
M6	Equal variances assumed	.114	.737	13.656	38	.000	.44005	.03222	.37482	.50528
	Equal variances not assumed			13.656	37.583	.000	.44005	.03222	.37479	.50531
M7	Equal variances assumed	1.067	.308	17.916	38	.000	.53435	.02982	.47397	.59473
	Equal variances not assumed			17.916	36.471	.000	.53435	.02982	.47389	.59481
M8	Equal variances assumed	10.333	.003	10.599	38	.000	.47450	.04477	.38387	.56513
	Equal variances not assumed			10.599	27.992	.000	.47450	.04477	.38279	.56621
M9	Equal variances assumed	.445	.509	14.453	38	.000	.38625	.02672	.33215	.44035
	Equal variances not assumed			14.453	37.331	.000	.38625	.02672	.33212	.44038
M10	Equal variances assumed	14.903	.000	22.963	38	.000	.47130	.02052	.42975	.51285
	Equal variances not assumed			22.963	23.849	.000	.47130	.02052	.42892	.51368
M11	Equal variances assumed	.349	.558	8.201	38	.000	.40365	.04922	.30401	.50329
	Equal variances not assumed			8.201	37.540	.000	.40365	.04922	.30397	.50333
M12	Equal variances assumed	.032	.859	7.933	38	.000	.41495	.05231	.30906	.52084
	Equal variances not assumed			7.933	37.954	.000	.41495	.05231	.30906	.52084

The Levene's test that is beneficial to test the equality of variances is conducted for the dataset under consideration. The test is conducted for 20 cluster centers in the two categories of fake (F) and authentic profiles (A). The test clearly illustrates that some of the metrics are insignificant in computing the overall outcome. The resulting p-value for these insignificant metrics is greater than 0.05 which is usually taken as the cutoff for significance. The metrics Status Count (M9), Emotion Stability (M1), @ Mentions (M7), Friends Count (M11), Unique Words (M6) and Followers Count (M12) depict insignificance.

However, the two-tailed t-test for equal means reflects significance of all the given metrics which indicates that these metrics are significantly different for the two groups. This is indicative of that fact that even though the dispersion of cluster centers across the two groups of profiles is similar, the cluster centers are actually different data points. This can be firmly asserted since the means of the centroids show significant difference across the entire set of metrics. The cluster centers for the entire set of instances is subsequently modeled to find the respective contribution to the five personality dimensions adopted from the big five framework. The overall centroids thus obtained from the analysis can be used by any supervised or unsupervised model for predictive analysis. Table 5.6 illustrates the overall cluster centroids for the two categories of Twitter users.

Table 5.6: Overall Cluster Center Identification using GWO-KNN and ABC-KNN

S. No.	Personality Dimension	Metric Name	GWO-kNN		ABC-kNN		Overall Centroids	
			F	A	F	A	F	A
1.	Extroversion	Status count (M9)	0.616	0.326	0.673	0.160	0.644	0.243
		Unique words (M6)	0.667	0.199	0.536	0.125	0.602	0.162
		HTML link count (M5)	0.765	0.129	0.682	0.134	0.724	0.132
2.	Agreeableness	Favorites count (M10)	0.602	0.107	0.512	0.094	0.557	0.100
		@ Mentions (M7)	0.750	0.332	0.873	0.196	0.812	0.264
3.	Conscientiousness	Friends count (M11)	0.881	0.535	0.695	0.211	0.788	0.373
		Hashtag freq. (M3)	0.857	0.093	0.954	0.043	0.906	0.068
		Follower Count (M12)	0.914	0.500	0.464	0.081	0.689	0.290
4.	Neuroticism	Emotion stability (M1)	0.485	0.384	0.875	0.055	0.680	0.220
		Polarity stability (M2)	0.906	0.050	0.931	0.067	0.918	0.058
5.	Openness to Experience	Unique hash tags (M4)	0.872	0.072	0.808	0.069	0.840	0.071
		Lexical diversity (M8)	0.611	0.323	0.691	0.100	0.651	0.211

Further, the findings can be validated by comparison with the existing fake profile detection approaches and the parent approaches. Table 5.7 depicts the results with GWO-KNN outperforming ABC-KNN in terms of convergence which proved to be a disadvantage as it leads to premature convergence as discussed in literature and reduces the accuracy in return.

Table 5.7: Comparative Analysis of Results for Fake Profile Detection

Approach	Algorithm	Avg. Time (seconds)	Convergence Iterations	Accuracy %
Existing [157]	SMO-Poly Kernel	N/A	N/A	68.47
	J48	N/A	N/A	65.81
	SMO-Normalized Poly Kernal	N/A	N/A	65.29
	Random Forest	N/A	N/A	59.79
	kNN k=10	N/A	N/A	59.7
	kNN k=3	N/A	N/A	59.39
	kNN k=5	N/A	N/A	33.91
	Naive Bayes	N/A	N/A	61.06
Parent	GWO [26]	87.56	**2**	**80.17**
	ABC [23]	96.78	4	83.28
	kNN [158]	120.23	4	**80.23**
Proposed	ABC-kNN	98.83	4	**87.09**
	GWO-kNN	89.65	**2**	83.04

It is evident from the comparison that ABC-KNN outperforms GWO-KNN and the remaining heuristic approaches in terms of accuracy. This is both due to the relevant metrics take for modeling the outliers and the hybrid approach that avoids local optimum while converging to the outliers. The GWO-KNN on the other hand converges in the least number of iterations.

5.2.5 Conclusion and Future Scope

In the current scenario, various social media platforms are being used in diverse ways for the promotion of online content targeting a large group of people. This creates a race for traction and often results in the generation of fake profiles that help in artificial propagation of content over the web. Studies in existing literature focus on detecting spam and identification of these fake profiles specifically on platforms like Twitter where opinions change views of the masses. This work uses a set of attributes to predict whether a profile is "fake" or not. The metrics used are diverse in nature and try to capture different aspects of a user's personality using a personality framework. There has been no evidence in literature surrounding the use of these descriptive and content metrics mined from Twitter and mapped to personality dimensions.

This work focuses on proposing a set of metrics for detection of fake Twitter profiles. The metrics have been logically grouped into five personality traits and thus it is the first attempt for identification and analysis of outliers in this trending domain using personality dimensions. The work uses a comprehensive set of 10,000 and 555,684 tweets for over 12 finalized metrics. Two hybrid bio-inspired approaches GWO-KNN and ABC-KNN have been used for the purpose of mining outliers. Methodologically, this work tries to overcome the limitation of locally optimum solutions that often occurs in traditional approaches with the use of bio-inspired computing algorithms which act as black boxes for the analysis making the approach simple and flexible to apply on any dataset.

The detection of fake Twitter profiles can be useful in multiple domains including marketing [159], governance [160], spam detection and control [161]. Marketers can use these findings for the identification of potential influencers that can be targeted for gaining higher traction on their post. The findings can provide gainful insights for viral marketing where the content propagates through word of mouth largely depending the user engagement on social media. The current work can be further extended to include network dynamics including centrality analysis (degree and betweenness centrality) to gauge the effect of network ties in identification of these profiles.

CHAPTER 6

6. INTEGRATING CHAOS FOR OUTLIER DETECTION

This sections attempts to improvise on the convergence speed by exploring a better search space while optimizing through the uses of chaos theory. In almost all meta-heuristic algorithms with stochastic components, random behavior is obtained by using various probability distributions usually Gaussian. It can be advantageous to replace such random components with chaotic maps since they possess similar properties of randomness with better statistical and dynamical properties. The chapter introduces chaotic firefly and cuckoo search approaches for segregating popular online content and identifying spam in search engines.

6.1 Segregating popular online content

6.1.1 Introduction

The increased internet use has a great impact on the way business is done specifically when it comes to marketing in the B2B domain. The adoption of e-business further affects the final outcome of businesses in terms of their performance [66]. The prominent information and communication technology with the advancements in the internet have led to the emergence of Web 3.0 [3]. The growing importance of Web 3.0 including social media and online content promotion web portals has greatly impacted the way organizations manage market their products and services [2]. The content that is propagated and shared at a faster pace reaching a larger audience is often considered to be viral. Organizations are thus leveraging the power of social transmission and virality of their online content to promote goods and services [130]. Further, existing studies investigate the factors that might affect the virality of online content. The studies also discuss about how virality may be explored in Web 3.0 domains including social media and micro-blogging websites. These content publishing websites and social media platforms apart from catalyzing information propagation also generate large amount of user

generated content (UGC) [10]. The discussions and the UGC on the web is further indicative of how consumers perceive and opine. This consumer perception and belief has impacted businesses to a great extent. The UGC may also affect the popularity of content on the web amongst other factors and thus becomes an interesting area to explore. This study thus uses several metrics including the analysis of UGC for sentiment, content and descriptive analytics [129] for predicting what content might become popular.

The current study hence proposes a chaotic cuckoo search algorithm [162] that is used for predicting the popularity of online content. The algorithm is further integrated with k-means to optimally cluster the content into popular and not so popular categories based on the identified metrics [119]. The proposed approach can be scaled for large amount of textual content from any domain. Findings may be useful in domains of digital marketing, e-commerce, social media marketing and e-governance amongst others to predict potential content that has chances of becoming popular and subsequently viral. This may be useful for the promotion of new launches in the service and retail industry [163]. The subsequent sections discuss the review of existing literature, the research methodology which includes the data description highlighting the metrics considered for analysis followed by the proposed approach. Finally, the results and analysis section compares the proposed approach for twelve chaotic maps.

6.1.2 Background

Organizations now are eyeing for user attention, and the visibility of these firms on the web has associated business value that may affect the firm's growth. Studies in existing literature have shifted their focus towards e-businesses, developing theoretical foundations for value creation [69]. The adoption of the e-business strategies in the current scenario further impacts the business performance of firms [66] specifically when it comes to adoption of digital marketing services in Web 3.0 domain [134]. Content popularity in the form of effective word of mouth thus becomes the major driver to marketing in the digital era.

It therefore becomes critically essential for organizations to know what factors drive content popularity. The existing literature does not have a lot of studies that discuss the factors affecting the popularity of content. Szabo and Huberman [164] predict the long term popularity of content on Digg, Youtube and Vimeo with user's access metrics. Further, the popularity of published articles on web portals and video popularity has been examined by using UGC [165] [166]. The popularity of articles on Wikipedia has also been investigated by considering metrics like traffic on the article, number of clicks to the article by users and the hyperlinks pointing to the same [167].

The advent of Web 3.0 has further changed the scenario with social media platforms like Twitter and Facebook coming into the picture. The focus is primarily on how popularity of social media posts affects the social media marketing strategy of firms. Existing studies use machine learning approaches for modeling the same [168]. But there are no studies that utilize meta-heuristics to optimize the location of potential solutions. This study however, uses some metrics from social media in terms of social shares along with other metrics from the content to cluster and predict the popularity. The subsequent sub section focuses on the existing literature of bio inspired computing algorithms that are used for the prediction.

This study uses one of the most popular bio inspired approach which mimics the behavior of cuckoo birds [22]. The algorithm is known to produce promising results for optimizing engineering and structural problems [169] [171]. The recently proposed variant chaotic cuckoo search (CCS) variant is known to converge to a globally optimal solution faster. It further demonstrates better accuracy when evaluated on benchmarked objective functions when compared to existing approaches [163]. The work proposes a hybrid chaotic cuckoo search integrated with k-means [119] for identifying popular and not so popular online content.

6.1.3 Methodology

This section focuses on the methodology adopted for identification of popular online content. The details of the metrics used for the analysis are described in Section 3.5. The metrics are used to model and cluster the articles into popular vs. not popular. Further, the total shares metric is used for the purpose of validation of the proposed approach. The subsequent section demonstrates the k-means integrated chaotic cuckoo search approach.

Hybrid Chaotic Cuckoo Search Algorithm

The cuckoo search (CS) algorithm mimics the brooding behavior of cuckoos. The movement of cuckoos towards nests is done via Levy flights. The CS algorithm is primarily a population based search algorithm for finding a globally optimal solution. The cuckoos lay their eggs in the nest of the other birds referred to as host birds. Further, cuckoos also imitate the patterns of host birds to avoid abandoning their eggs by them. The cuckoo eggs in the nests are representative of a potential solution [22]. A new solution is obtained by means of Levy Flights. The new solution (pos_i^{t+1}) is obtained using the previous candidate solution (pos_i^t) in combination with Levy ($Levy(\delta)$).

$$pos_i^{t+1} = pos_i^t + Step\ Size \oplus Levy(\delta) \tag{6.1}$$

where, $s > 0$

and $Levy \sim u = v^{-\delta}$ (6.2)

The step size in this case is updated using the chaotic maps. The current study considers a set of **twelve chaotic maps** for the analysis [169] including circle, intermittency, Gaussian, tent, sinusoidal, chebyshev, piecewise, logistic, sine, liebovitch, singer and iterative map. The initial step size is computed using the Mantegna's algorithm for a stable Levy flight [170]

$$Step\ Size = \frac{u}{|v|^{1/\delta}} \tag{6.3}$$

The values $u \sim NormDist(0, {\sigma_u}^2)$ and $v \sim NormDist(0, {\sigma_v}^2)$, where $\sigma_u =$

$$\left[\frac{Gamma(1+\delta)\sin\left(\frac{\pi\delta}{2}\right)}{Gamma\left[\frac{(1+\delta)}{2}\right]\delta 2^{\frac{\delta-1}{2}}}\right]^{\frac{1}{\delta}} \tag{6.4}$$

and $\sigma_v = 1$, $\delta = 3/2$ describe the Levy flight.

The chaotic cuckoo search approach is integrated with k-means for clustering [119]. The pseudo-code for the proposed approach is depicted in Figure 6.1.

Begin
Step 1: Initialization. *Set the iteration counter t = 1.*
Initialize the population of host nest P randomly;
Set discovery rate (d) and initial value of the chaotic map randomly.
Determine number of clusters, K.
Assign K clusters for each host nest P
Assign K objects as initial centroids for each nest.
Step 2: Exploration. *Calculate fitness of centroid for each cluster*
While *t < Max* ***do***
Sort the population according to their fitness.
Update the step size using chaotic maps
Select a cuckoo i randomly
Replace the corresponding solution by Lévy flight (Equation 1-4).
Evaluate its fitness Fit_i.
Identify a nest j randomly.
if ($Fit_i < Fit_j$)
Replace j by the newly obtained solution.
end if
Abandon fraction (d) poor nests and build new nests.
Step 3: Centroid Update. *Sort the population and find the current best.*
t = t+1.
Reassign the cluster centers
Output best cluster configuration based on best fitness value.
end while
End.

Figure 6.1: Pseudo-code of k-Means integrated Chaotic Cuckoo Search (CCS)

The study uses the twelve chaotic maps for k-means integrated chaotic cuckoo search. The results are further compared with k-means integrated with original cuckoo search approach. The subsequent section discusses the results for the proposed approach.

6.1.4 Analysis and Findings

The current study proposes the CCS approach integrated with k-means for identifying the cluster of popular and not popular online content based on a set of significant metrics obtained after a statistical analysis. The use of chaotic maps improvises the search for a globally optimum solution in the search space. The description of all chaotic maps used in the work is illustrated in Figure 6.2. Along with the quality of the solution, the convergence to the solution is also faster.

The 19 metrics obtained after the statistical t-test are used to model and cluster the news articles into popular and non-popular using the proposed approach. For the purpose of validation of the results, the dependent variable in terms of Total shares is considered. The dataset uses a threshold of 1400 for segregating the popular and not popular Mashable news articles. The same has been used for validation of the obtained results after normalizing the output variable using a similar min-max normalization method as used for the independent defining metrics

The proposed approach with Singer chaotic map (as described in (6.5)) gives the highest accuracy as compared to the remaining variants.

$$x_{i+1} = \upsilon(7.86x_i - 23.3x_i^2 + 28.7x_i^3 - 13.3\,x_i^4) \tag{6.5}$$

where $\upsilon\epsilon$ [0.9,1.08].

No.	Name	Definition
M1	Chebyshev map	$x_{k+1} = \cos(k \cos^{-1}(x_k))$
M2	Circle map[a]	$x_{k+1} = x_k + b - (a/2\pi)\sin(2\pi k) \bmod(1)$
M3	Gaussian map	$x_{k+1} = \begin{cases} 0 & x_k = 0 \\ 1/x_k \bmod(1) & \text{otherwise} \end{cases}$, $1/x_k \bmod(1) = \frac{1}{x_k} - \left[\frac{1}{x_k}\right]$
M4	Intermittency map	$x_{k+1} = \begin{cases} \varepsilon + x_k + c x_k^n & 0 < x_k \le P \\ \frac{x_k - P}{1-P} & P < x_k < 1 \end{cases}$
M5	Iterative map	$x_{k+1} = \sin\left(\frac{a\pi}{x_k}\right), \quad a \in (0, 1)$
M6	Liebovitch map	$x_{k+1} = \begin{cases} \alpha x_k & 0 < x_k \le P \\ \frac{P - x_k}{P_2 - P_1} & P_1 < x_k \le P_2 \\ 1 - \beta(1 - x_k) & P_2 < x_k \le 1 \end{cases}$,
M7	Logistic map	$x_{k+1} = a x_k (1 - x_k)$
M8	Piecewise map	$x_{k+1} = \begin{cases} \frac{x_k}{P} & 0 \le x_k < P \\ \frac{x_k - P}{0.5 - P} & P \le x_k < \frac{1}{2} \\ \frac{1 - P - x_k}{0.5 - P} & \frac{1}{2} \le x_k < 1 - P \\ \frac{1 - x_k}{P} & 1 - P \le x_k < 1 \end{cases}$
M9	Sine map	$x_{k+1} = \frac{a}{4}\sin(\pi x_k), \quad 0 < a \le 4$
M10	Singer map	$x_{k+1} = \mu(7.86 x_k - 23.31 x_k^2 + 28.75 x_k^3 - 13.302875 x_k^4)$
M11	Sinusoidal map	$x_{k+1} = a x_k^2 \sin(\pi x_k)$
M12	Tent map	$x_{k+1} = \begin{cases} \frac{x_k}{0.7} & x_k < 0.7 \\ \frac{10}{3} & x_k \ge 0.7 \end{cases}$

[a] With $a = 0.5$ and $b = 0.2$, it generates chaotic sequence in (0, 1)

Figure 6.2: Description of Chaotic Maps

(Adopted from [163])

Figure 6.3 demonstrates the clustering plots obtained from the CCS with Singer Map for the same with popular news content in green and non-popular news content highlighted in red. The clusters have overlap and are not crisp because the problem definition is subjective in terms of how popularity can be gauged collectively using the selected metric. Figure 6.4 presents the convergence plot for the same and it is evident that it converges to a solution in the last iteration.

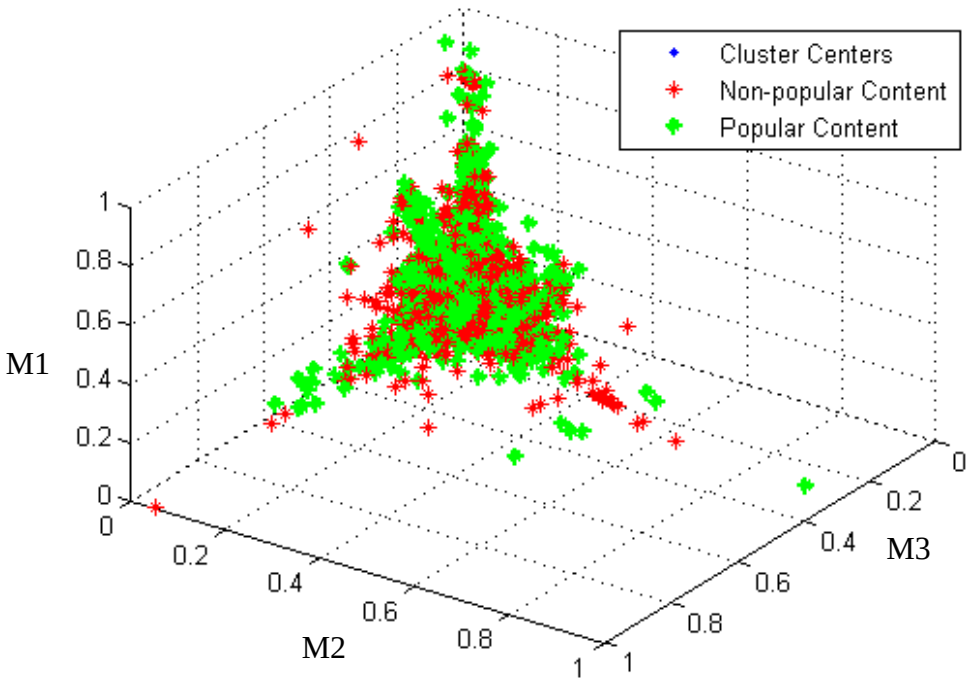


Figure 6.3: Identification of Popular Content using CCS (Singer Map)

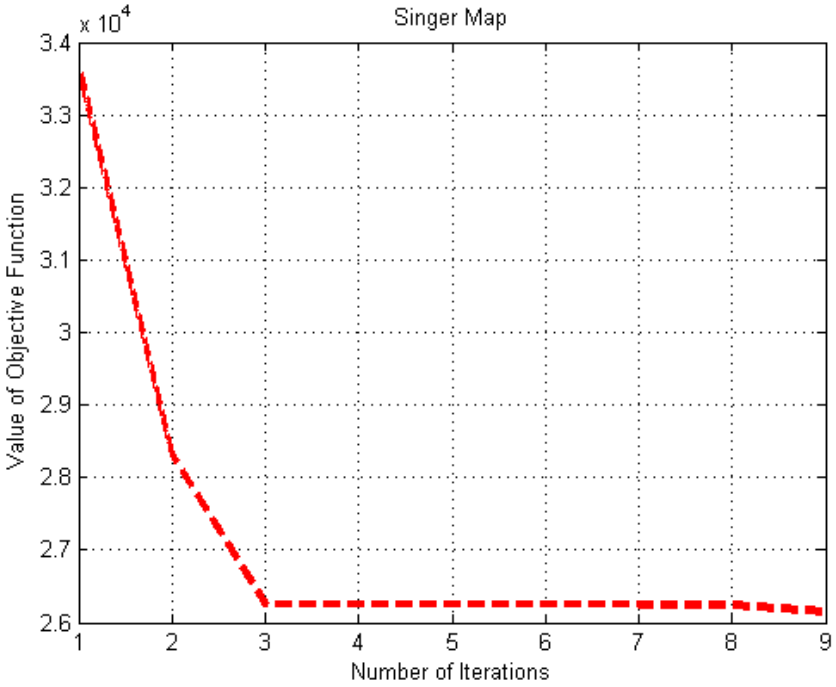


Figure 6.4: Convergence Plot for Singer Map

On comparing the convergence speed, Figure 6.6 illustrates the convergence plots for the twelve chaotic map variants for CCS along with the original CS approach without chaos theory to update the step size.

The Sine map as depicted in (6.6) produces the best results in terms of convergence and reaches to a globally optimal solution over minimum iterations.

$$x_{i+1} = \frac{\alpha}{4}\sin(\pi x_i) \tag{6.6}$$

where $1 \leqq \alpha \leqq 4$, for the purpose of this study the value of α is taken as 4.

The convergence plot for the Sine Map variant is depicted in Figure 6.5 and is known to converge to an optimal solution in the second iteration itself.

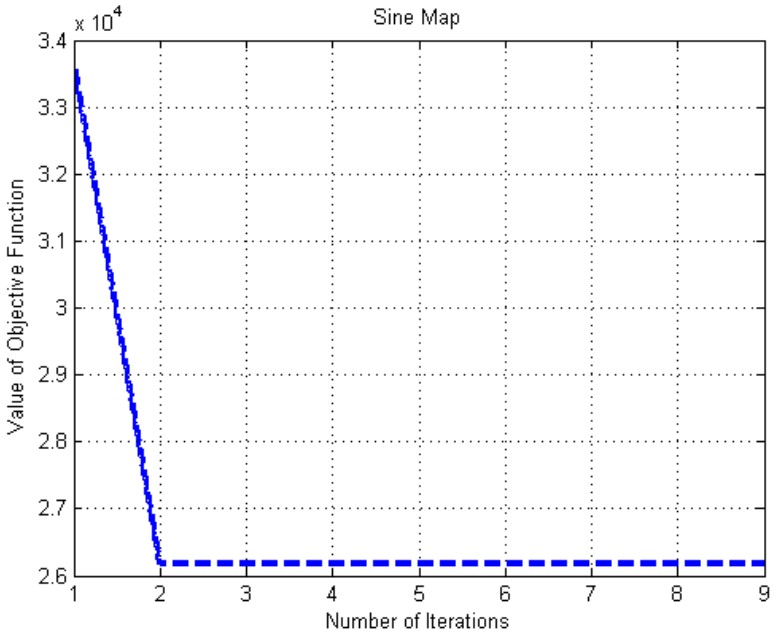


Figure 6.5: Convergence Plot for Sine Map

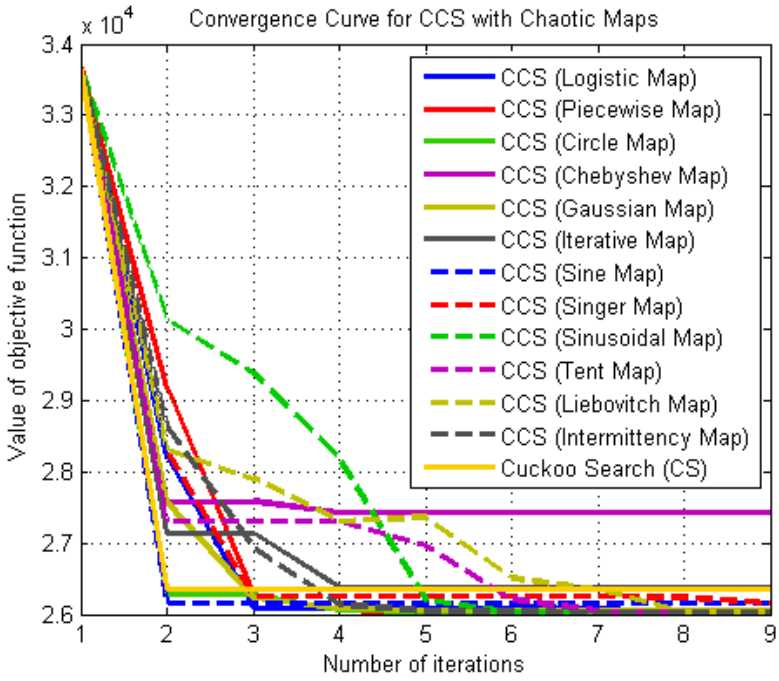


Figure 6.6: Comparative Convergence Curves for CCS with Chaotic Maps

The convergence of the Sine map CCS variant and the CS without chaos is same while on comparing the accuracy, the CCS Sine and Singer map both outperform CS without chaos with respective accuracies of 94.23% (CCS Sine Map), 96.78% (Singer Map) and 93.78% (CS without chaos). Further, the running time for each of the 12 variants of k-means integrated CCS and CS (without chaos) are illustrated in Table 6.1. This is the average running time computed over 10 pseudo steps. The sinusoidal map takes the least time (217.2 seconds) for computing the clusters. However, it does not converge to an optimal solution in that time. The Gaussian map on the other hand takes the longest time (403.9 seconds).

Table 6.1: Running Time of Proposed CCS Variants

S. No.	Running Time for Chaotic Cuckoo Search variants	
	Variant	**Time elapsed (seconds)**
1.	Circle Map	282.6
2.	Intermittency Map	276.9

S. No.	Running Time for Chaotic Cuckoo Search variants	
	Variant	**Time elapsed (seconds)**
3.	Gaussian Map	403.9
4.	Tent Map	242.3
5.	Sinusoidal Map	217.2
6.	Chebyshev Map	280.7
7.	Piecewise Map	285.9
8.	Logistic Map	258.8
9.	Liebovitch Map	245.1
10.	Singer Map	274.1
11.	Iterative Map	320.7
12.	Sine Map	254.9
13.	CS without Chaos	239.1

The cluster centers are computed for the best CCS variants including Sine and Singer map along with CS without chaos theory, the same are depicted in Table 6.2.

For each of the chaotic map integrated with k-Means CCS, the cluster center for popular (P) and not so popular (N) content is demonstrated. These cluster centers can be used for testing new news articles with the same metrics to categorize them into the two sets. As seen in the cluster plot, the cluster centroids further highlight the not so crisp distinction between the centroids for the two clusters. This opens new avenues of exploring the fuzzy overlap in the two identified categories in the future studies.

Table 6.2: Cluster Centers Computed using K-Means Integrated CCS

Approach		Cluster Centers																		
		MI1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**M9**	**MI10**	**MI1**	**MI2**	**MI3**	**MI4**	**MI5**	**MI6**	**MI7**	**MI8**	**MI9**
Singer Map	P	0.21	0.07	0.15	0.06	0.75	0.02	0.05	0.08	0.03	0.87	0.04	0.54	0.47	0.43	0.65	0.68	0.75	0.42	0.62
	N	0.04	0.03	0.04	0.01	0.69	0.01	0.20	0.15	0.24	0.24	0.25	0.44	0.46	0.35	0.74	0.28	0.54	0.68	0.16
Sine Map	P	0.03	0.03	0.05	0.01	0.65	0.01	0.06	0.82	0.05	0.09	0.06	0.46	0.45	0.36	0.70	0.31	0.53	0.61	0.16
	N	0.04	0.03	0.03	0.01	0.70	0.01	0.21	0.10	0.25	0.25	0.27	0.44	0.46	0.35	0.74	0.28	0.54	0.69	0.16
CS No Chaos	P	0.03	0.03	0.03	0.01	0.65	0.01	0.30	0.17	0.39	0.10	0.13	0.43	0.45	0.34	0.74	0.38	0.55	0.51	0.21
	N	0.04	0.03	0.04	0.02	0.72	0.01	0.12	0.14	0.12	0.35	0.35	0.45	0.46	0.36	0.74	0.21	0.52	0.82	0.11

Further, the proposed hybrid bio-inspired approaches are compared with existing traditional machine learning approaches in literature in terms of accuracy, precision, recall and the F-measure. The chaotic variants outperform the same along with the non-chaotic cuckoo search approach. The results for the same are highlighted in Table 6.3.

Table 6.3: Comparative Analysis of Results for Content Popularity

Approach	Model	Accuracy	Precision	Recall	F-Measure
Existing [168]	Random Forest (RF)	0.67	0.67	0.71	0.69
	Adaptive Boosting (AdaBoost)	0.66	0.68	0.67	0.67
	Support Vector Machine (SVM)	0.66	0.67	0.68	0.68
	K-Nearest Neighbors (KNN)	0.62	0.66	0.55	0.60
	Naïve Bayes (NB)	0.62	0.68	0.49	0.57
Proposed [169]	Cuckoo Search (w/o Chaos)	0.93	0.93	0.87	0.89
	Cuckoo Search (Singer Map)	**0.96**	**0.94**	0.89	0.91
	Cuckoo Search (Sine Map)	0.94	0.93	**0.92**	**0.92**

6.1.5 Conclusion and Future Research Directions

The work attempts to identify popular content in online news. A set of 39,797 news articles modeled using 19 statistically significant metrics. The contribution is both in the domain of content popularity and in terms of methodology. The metrics include link count, Mashable link count, image, video and keyword count, Mashable article share, content/title subjectivity and

sentiment, avg. positive and negative polarity amongst others. The study proposes a k-means integrated chaotic cuckoo search approach for clustering and identifying popular news content.

Findings indicate that the Singer map variant of the approach outperforms the remaining in terms of accuracy (96.78%) while the Sine map expedites the search for a potential globally optimal solution faster achieving the convergence fastest. The remaining ten chaotic maps however outperform the original cuckoo search approach when integrated with k-means in terms of accuracy but converge to a solution slower. Future studies may focus on exploring the overlap using fuzzy clustering approaches for content that has equal probability of getting popular based on the selected metrics. Findings of the study have applicability in domains of e-commerce, e-governance, social media and influencer marketing to predict what content may become popular and subsequently viral.

6.2 Identifying Spam websites in Search Engines

6.2.1 Introduction

With the advent of interactive web and increased use of the internet by people, the visibility of content on the web is of prime importance. Both profit and non-profit firms nowadays are majorly interested in online marketing of their ideas, services, products and projects. These organizations have realized that the web plays an inevitable role and attracts significant marketing opportunities. Literature highlights the important factors for online marketing and their relevance in making the content popular among the users [85]. Now, a lot of traffic to these organization websites comes from organic search queries in search engines like Google [86], Bing and Yahoo to name a few. Literature highlights the use of SEM in various domains including tourism, e-commerce and marketing [91] [92] [93]. It is noticed that people generally tend to use the top results from their search query, making the rank of web pages of vital importance for the companies. As a result, organizations have started adopting strategies for brand positioning using SEM [94].

An important technique used by companies to improve the ranking of their pages is search engine optimization (SEO) a major tool in SEM that plays a critical role in increasing the web pages' visitor count. This is usually done by ranking it higher on the search results often using keywords that describe the website's content [95]. SEO techniques are thus critical in increasing the visibility of a website on the internet and attract greater organic search traffic [96]. SEO techniques used in business are often black hat in nature and do not lie within the SEO guidelines and often result in poor ranking and blacklisting of the website from the search engine when detected. These techniques comprise of cloaking, use of doorway pages and invisible elements [97].

Further, SEO can be categorized into off-page and on-page optimization; Off-page optimization focuses on link purchases and link building, while on-page optimization focuses on high quality content and its presentation including website structure, multimedia, keyword management,

accessibility and portability. The focus of this study would be on the former majorly link building, search engines often prioritize pages based on the number of back links to it [110]. The process of link building is known to improve the SERP/page rank of the page on the search engine. This has resulted in paid link building services that are not deemed favorable by search engines. However, these paid listings are acceptable in website listing hubs like Moz.

The websites that are used for link building may not always be authentic and thus include a lot of spam, not trustworthy web pages. These websites therefore try to increase the content to build more and more links surrounding keywords. The content is often not original and manipulated. Techniques like article spinning, link farming and keyword stuffing are popular for recreating, manipulating and building content for link building. These are often not beneficial for business and Google has introduced several changes and algorithms to weed out such spam websites in the past.

6.2.2 Background

There are very few discussions in academic literature surrounding the detection of these black hat SEO link building approaches [111] [172] [173]. The purpose of this study is to thus focus on mining outlier websites that use these black hat SEO practices including link building, article spinning, link farming and keyword stuffing to generate content. Relevant metrics have been identified for two different cases for analysis and detection of such spam profiles. Further, bio inspired algorithms have been used to classify the websites as trustworthy or spam based on the identified metrics. Metric analysis for selection of websites using cuckoo search is also available in literature [174].

The focus of this work is primarily on identifying websites for link purchases, however while selecting such websites it is often seen that majority of them resort to Black Hat SEO including article spinning, link building and keyword stuffing. Back links purchased from such websites would have an adverse effect on the client opting for SEO for an enhanced SERP. This section of work thus attempts to develop a mechanism for detecting such spam untrustworthy websites

that shall be avoided while investing in link purchases. For the purpose of detecting such outlier websites a mix of publically available metrics derived by off-site analytics has been modeled by using bio inspired computing algorithms. This would create a mechanism for website assessment so that potentially harmful websites may be filtered out while selecting web pages for SEO. This would help avoiding selection and investment on non-trustworthy websites for subsequent search engine marketing and penalization from search engines.

There are no significant discussions surrounding these metrics in academic literature and majorly various link data providers like Moz , Majestic, Ahref and Webmaster tools have developed ranking mechanisms that are used worldwide for analyzing the position of a page in SERP. Since the rank of a web page is an important indicator of its visibility to users. Organizations resort to techniques that allow them greater visibility on the web when the search queries are made on various search engines. Search engine optimization (SEO) thus comes into picture and plays a vital role in improvising the rank of web pages in SERP.

6.2.3 Methodology

The ranking of websites on search engines is of prime importance to have a comparative analysis of the same. As discussed there are several metrics for computing a website's rank. These metrics have been given by different companies and comprise of over lapping criteria that suit their individual needs. For the purpose of this study we are considering two separate cases for different organizations. The dataset used for the analysis of this study thus comprised of two different cases: Case 1 and Case 2 with 1070 data points and 1682 data points respectively with each data point belonged to a separate website. The details of which are mentioned in the Section 3.6. The two cases had web pages belonging to separate domain areas and categories, thus Delphi was separately employed for the two studies. Case 1 is for a knowledge portal initiated in 2009 called Business Fundas (http://business-fundas.com) comprising of articles related to managerial subjects like supply chain, strategy, marketing, finance and e-commerce to name a few. Case 2 is for Tech Talk (https://tech-talk.org), a knowledge portal initiated in 2012 focusing

on articles related to domains of information technology like big data analytics, e-commerce, business analytics, social media and related domains.

For Case 1, a total of eight practitioners were a part of the Delphi process, the consensus was reached in three iterations. The experts comprised of Business Funda's CTO, CEO, business development head, marketing heads and employees working in the organization since inception. The experts have an experience of 8 to 12 years in the domain with an average experience of 9.4 years. All eight experts recorded a score between 3 to 4 having a median of above 3.25, making it sufficient to reach the consensus. The consensus was achieved in three iterations based on the requirements. On the contrary, the Case 2 required two iterations to reach the consensus with the panel comprising of six experts with Tech Talk's CTO, CEO, business developments and marketing heads having an average experience of 6.1 years. The practitioners' recorded a median lying between 3.5 and 4.0. As per Green [177] if 80% of the panelists, allot a score between 3 and 4 with a median of 3.25 and above, the consensus is said to be achieved. Table 6.4 shows the iteration scores for the two cases with mean ($\bar{\mathbf{x}}$) and median ($\tilde{\mathbf{x}}$) values for every metric.

Table 6.4: Iteration Scores for Metric Identification using Delphi

Initial Metrics	**Case 1-Business-Fundas.com**						**Case 2 – Tech-Talk.org**			
	Iteration Scores						**Iteration Scores**			
	I		**II**		**III**		**I**		**II**	
	$\bar{\mathbf{x}}$	$\tilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\tilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\tilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\tilde{\mathbf{x}}$	$\bar{\mathbf{x}}$	$\tilde{\mathbf{x}}$
Page Rank	3.25	3.00	3.38	3.50	3.50	3.50*	3.25	3.00	3.50	3.50*
Page Authority	2.63	2.50	2.75	3.00	2.63	3.00	3.00	3.00	3.38	3.50*
Domain Authority	3.38	3.00	3.50	3.50	3.50	3.50*	3.38	3.00	3.50	3.50*
MozRank	1.63	1.50	1.38	1.00	1.38	1.00	1.50	1.50	1.50	1.50
MozTrust	1.63	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
Citation Flow	2.38	2.50	2.50	2.50	2.50	2.50	2.38	2.50	2.00	2.00
Trust Flow	2.50	2.50	2.63	2.50	2.63	2.50	2.50	2.50	2.25	2.00

Alexa Rank	2.88	3.00	3.13	3.00	3.25	3.50*	3.25	3.50	3.63	4.00*
URL Rating	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
Domain Rating	1.63	1.50	1.50	1.50	1.50	1.50	1.63	2.00	1.50	1.50
Ahref Rank	1.50	1.50	1.38	1.00	1.38	1.00	1.63	1.50	1.38	1.00
Back Links	1.75	2.00	2.00	2.00	1.88	2.00	2.25	2.50	1.88	2.00
Total Links	2.00	2.00	2.25	2.00	2.00	2.00	2.13	2.00	1.63	1.50
Google Index	2.50	2.50	2.50	2.50	2.38	2.50	2.75	3.00	3.50	3.50*
Social shares	3.38	3.50	3.50	3.50	3.50	3.50*	3.50	3.50	3.38	3.50*
Domain Age	3.50	3.50	3.50	3.50	3.63	4.00*	3.50	3.50	3.63	4.00*

*The highlighted entries are for the finalized metrics after Delphi consensus for the two case studies

The final metrics achieved after the Delphi Technique for Case 1 (Business Fundas) comprises of Page Rank (M1), Domain Authority (M2), Alexa Rank (M3), Social shares (M4) and Domain Age (M5). The metrics for Case 2 (Tech Talk) includes Page Rank (M1), Page Authority (M2), Domain Authority (M3), Alexa Rank (M4), Google Index (M5), Social shares (M6) and Domain Age (M7). Thus, the final datasets after the Delphi study comprised of 5 metrics associated with for 1070 websites for Case 1 and 7 metrics for 1682 websites of Case 2. Since, these metrics had varied range min-max normalization was used to scale the parameters on a 0-1 scale before outlier detection could be used to mine spam websites.

The purpose of the study is to identify relevant metrics for website assessment and subsequent outlier detection to segregate spam websites. The K- Means approach for clustering works well in the sense that it has an ability to find good cluster centroids at the start but the only pitfall is, it may result into a local optima. The K-means approach may be divided into four phases namely: Initialization, Cluster assignment, Centroid Updation and finally the Evaluation or Exploration phase. Bio inspired optimization approaches are considered to solve the above mentioned drawbacks over iterations to produce a globally optimum solution. Tang et al. [119] have modified the K-Means approach by including an optimization phase which optimizes the

solution with subsequent iterations and replaces the current best solution with the new solution. The process of optimization is repeated until the stopping criterion is met.

This work focuses on modifying the K- Means firefly algorithm (FA) and compares it with the integrated bat and cuckoo search approaches. There are several variants of FA in the existing literature [179], specifically with chaos [174] and Levy. The FA is also known to produce promising results when used for clustering applications [180]. This study proposes a chaotic firefly algorithm integrated with K-Means for tuning two of the existing firefly coefficients to speed up the search in unknown environments. The hybrid chaotic optimization approaches are known to produce more efficient results and are often governed by statistical properties of the chaos sequences and the location of global optima [27].

Further, the exponential growth in the amount of structured and unstructured information generates a need of faster computation abilities and the chaos theory may help in achieving the optimization results faster. The study further does a comparative analysis of the proposed chaotic firefly approach with bat algorithm (BA) and cuckoo search algorithm (CS). Firefly algorithm mimics the patterns, characteristics and behaviors of fireflies; the light intensity of fireflies governs the best solution [24]. Bat algorithm works on echolocation behavior of bats, the loudness value is a criterion for identifying the best solution over iterations [25]. Lastly, the cuckoo search algorithm is based on the egg hatching patterns of cuckoo birds where the eggs represent potential solution [169] [171].

The methodology comprises of several steps including the initial cluster initialization, followed by optimization of initial cluster centers using the bio inspired approaches and subsequent clustering of the training data (60%) for the two cases into two clusters of outlier and authentic websites for SEM. Subsequently he remaining 40% of the data is classified using k-nearest neighbor approach after manual validation of the two clusters. The sub sections highlight the cluster initialization, bio inspired clustering and post data classification.

The K-Means clustering approach computes the initial cluster centers for the two clusters in this study by considering the significance in terms of belongingness to the cluster along with each website data point (*webdp*) and is given by:

$$significance_{i,K} = \begin{cases} 1, & webdp_i \in cluster_j \\ 0, & webdp_i \notin cluster_j \end{cases} \quad (6.7)$$

$$ClusterCenter_{K,M} = \frac{\sum_{i=1}^{SolSpace} significance_{i,K}\, webdp_{i,M}}{\sum_{i=1}^{SolSpace} significance_{i,K}}, \quad (6.8)$$

where *K=1…k*, *k* being the number of clusters, *SolSpace* refers to the solution space and *M=1…k*m, m* is the number of metrics under consideration.

The *m* value will be different for the two case studies under consideration, For Case-1 (Business Fundas), *m=5* whereas for Case-2 (Tech Talk), *m=7*. However, *k* remains the same as 2 for the two clusters of authentic and outlier websites. The subsequent distance between the cluster centers is given by:

$$F(ClusterCenter) = \sum_{K=1}^{k} \sum_{i=1}^{SolSpace} significance_{i,K} \sum_{M=1}^{k*m} \left(webdp_{i,M} - ClusterCenter_{K,M}\right)^2 \quad (6.9)$$

The initial cluster centroids are then updated over iterations by means of proposed chaotic firefly algorithm and with a minimization objective function depicted by:

$$F = \sum_{K=1}^{k} \sum_{i=1}^{Population} || \left| webdp_{i,K} - ClusterCenter_K \right| |^2 \quad (6.10)$$

The cluster initialization and subsequent clustering using the integrated approach is done for 60% data for the two cases, 1009 website data points for Case-1 and 642 data points for Case-2.

Chaotic Firefly Algorithm (FA)

The cluster centers achieved in the initialization phase are then updated iteratively to produce a globally optimum solution by means of bio-inspired computing approach. The firefly algorithm (FA) mimics the behavior of fireflies that move towards each other based on their brightness. The brightness of each firefly is given by:

$$\rho = \rho_0 \, e^{-\mu dist} \tag{6.11}$$

ρ_0 is the original intensity emitted by the firefly and μ is the absorption coefficient of the medium. Thus, related attractiveness (A) is defined as:

$$A = A_0 \, e^{-\mu dist^2} \tag{6.12}$$

A_0 being the attractiveness at distance (dist)= 0. The $dist_{ij}$ is the distance between the two fireflies i and j at positions pos_i and pos_j respectively:

$$dist_{ij} = \left|\left|pos_i - pos_j\right|\right| = \sqrt{\sum_{k=1}^{n} (pos_{i,k} - pos_{j,k})^2} \tag{6.13}$$

where $pos_{i,k}$ depicts the k^{th} component of pos_i for i^{th} firefly. The firefly with lower brightness and subsequent attractiveness moves towards the brighter, more attractive firefly having a greater attractiveness coefficient (A). The updated position pos_i is obtained by:

$$pos_i = pos_i + A_0 e^{-\mu dist_{ij}^2}(pos_j - pos_i) + \omega\beta_i, \; A_0 = 1, \omega \in [0, 1], \text{and } \mu = 1 \tag{6.14}$$

The updated position depends on the attractiveness coefficient (A) along with a randomization coefficient ω with a random variable vector β_i using a Gaussian distribution.

We further use chaotic maps to speed up the search for best solution in the environment. The chaos theory can be used in two ways in the firefly algorithm [175] which can be used to tune the absorption coefficient μ and the attractiveness coefficient A. It is observed that Sinusoidal and Gauss map are known to give best results for the two mentioned scenarios respectively. The sinusoidal map for tuning absorption coefficient μ can be represented as:

$x_{i+1} = \tau x_i^2 \sin(\pi x_i)$, $\tau = 2.3$, $x_o = 0.7$ and i represents the i^{th} chaotic map. (6.15)

While, the Gauss map for tuning the attractiveness coefficient A is:

$$x_{i+1} = \begin{cases} 0 & x_i = 0 \\ 1/x_i \bmod(1) & x_i \neq 0 \end{cases}, \text{ where } 1/x_i \bmod(1) = 1/x_i - [1/x_i]$$

(6.16)

Once the best solutions are identified using the bio inspired approaches, it is considered as the cluster centroids for the K-Means integration. Further, k-nearest neighbors approach is used to mine the outlier. Euclidean distance is used as the measure to compute the distance between each websites corresponding attributes and the best solution. The proposed chaotic firefly approach is compared with bat algorithm which is known to converge to an optimum solution the fastest when integrated with K-Means for identifying cluster centers and cuckoo search algorithm that is known to give the best results as per existing literature [119].

Further, there are existing studies surrounding cuckoo search algorithm for website selection [174] which has been used for benchmarking in this study. A comparative analysis of the three approaches thus used for detecting website outliers along with the parameter values taken are illustrated in Table 6.5.

Table 6.5: Pseudo-code for Chaotic FA, BA and CSA Algorithms

Chaotic Firefly Algorithm (FA)	**Bat Algorithm (BA)**	**Cuckoo Search (CS) Algorithm**
Begin Generate initial population for fireflies, $N_i, i = 1 \ldots . n$ Light intensity ρ_i for each N_i determined by *f(N)* Initialize light absorption coefficient μ For all fireflies in the population, If ($\rho_i < \rho_j$) Move *i* firefly towards *j*; $dist_{ij} = \left\|pos_i - pos_j\right\|$ is the Euclidean distance between the two fireflies at positions pos_i and pos_j The new position of the firefly, $pos_i += A\left(pos_j - pos_i\right) + \omega\beta_i$, that depends on β_i random variable vector and ω is the randomization coefficient. Attractiveness (*A*) can be varied $A = A_0\, e^{-\mu dist^2}$	Begin Generate initial population N_i and velocity, ω_i. Pulse frequency Pf_i at N_i is $Pf_i \in [0, Pf_{max}]$, Define Pulse Rate (*P*) Loudness (*L*) F: Random flight with velocity ω_i at position pos_i at time interval *t* $\omega_i^t = \omega_i^{t-1} + (pos_i^{t-1} - pos_{best}) * f_i$ and $pos_i^t = pos_i^{t-1} + \omega_i^t$ If (*F* > *P*) Select a solution among current best solutions (pos_{best}) If the (*F* < *L*) Accept new solution, reduce loudness $L_i^t = \tau L_i^{t-1}, \quad \tau \in (0,1)$ and increase pulse rate $P_i^t = P_i^{t=0}[1 - e^{-\phi t}]$, $\phi\ is\ constant > 0$. Keep the best solutions	Begin Generate initial population of 'n' host nests, $N_i, i = 1 \ldots . n$ Get a cuckoo i randomly by Levy Flights *L* Select a nest *j* among nests *n* Fitness of cuckoo *i*: *Fi* If (*Fi* > *Fi*) Replace *j* by new solution Abandon fraction of nests D_{rate}(worst nests) and build more solution pos_i at time *t* is the new solution by Levy *L*. $pos_i^{t+1} = pos_i^t + step \oplus L(\delta)$, *step* > *0*, $L \sim x = y^{-\delta}$ and $step = \frac{x}{\lvert y\rvert^{1/\delta}}$ $x \sim NormDist(0, {\sigma_x}^2)$ and $y \sim NormDist(0, {\sigma_y}^2)$ where $\sigma_x = \left[\frac{Gamma(1+\mu)\sin\left(\frac{\pi\mu}{2}\right)}{Gamma\left[\frac{(1+\mu)}{2}\right]\mu 2^{\frac{\mu-1}{2}}}\right]^{\frac{1}{\mu}}$ and $\sigma_y = 1$, where $\mu = 3/2$ (Mantegna, 1994) Keep the best solutions

which depends on brightness ρ: $\rho = \rho_0 e^{-\mu dist}$ Keep the best solutions Rank fireflies and find current best solution.	Rank bats and find current best solution.	Rank the nests and find the current best solution.
Parameter Values: ω (Randomness) :0.2 μ (Absorption Coefficient):1.0 Population: 400	**Parameter Values:** *L* (Loudness): 0.5 *P* (Pulse Rate): 0.5 Population: 400	**Parameter Values:** D_{rate} (Discovery Rate):0.25 Tol (Tolerance): $1.0e^{-5}$ Population: 400

After the optimum cluster centers are obtained the K-Means clustering is done to achieve the two cluster sets for authentic and outlier websites. Manual evaluation is used for validating these outlier websites. Subsequently the remaining 40% website data points are used for classification.

Outlier Website Classification

The previous sub sections discuss the initialization of initial cluster centroids for authentic and outlier websites and subsequent updation over iterations used the proposed chaotic firefly algorithm. This has been done along with a comparison with other popular bio inspired computing approaches integrated with K-Means that have been used in existing literature for the clustering and have shown promising results either in terms of accuracy or convergence speeds. The motive behind using the same was to achieve a globally optimum solution since K-Means being a simple and one of the most popular clustering approach often falls into a local optima. Further, chaotic maps are introduced in the proposed approach keeping in mind exponential growth in data both in structured and unstructured form making the approach computationally extensive. The chaos theory thus enables a faster convergence.

This sub section uses the k-nearest neighbor for the classifying the remaining 40% test data into the obtained clusters. A total of 673 data points are considered for Case-1 and 428 for Case-2. A 1-nearest neighbor is used to classify the websites to the nearest cluster using the Euclidean distance measure ($dist_{ij}$) given by $dist_{ij} = \left|\left|webdp_i - ClusterCenter_j\right|\right|$, where *i* represents the website data point and *j=1..k* for the two cluster centers. The outlier websites are thus classified into the cluster with the minimum distance.

After the optimum cluster centers are obtained the K-Means clustering is done to achieve the two cluster sets for authentic and outlier websites. Manual evaluation is used for validating these outlier websites. Subsequently the remaining 40% website data points are used for classification.

6.2.4 Analysis and Findings

This study focuses on detecting website outliers for the purpose of search engine marketing for the two cases: Business Fundas (Case 1) and Tech Talk (Case 2). The dataset considered for analysis comprised of 1070 data points (web pages) with 5 metrics namely Page Rank (M1), Domain Authority (M2), Alexa Rank (M3), Social shares (M4) and Domain Age (M5) for Case 1 and 1682 websites with a total of seven metrics including Page Rank (M1), Page Authority (M2), Domain Authority (M3), Alexa Rank (M4), Google Index (M5), Social shares (M6) and Domain Age (M7) for Case 2. The accuracy is in terms of both clustering using bio inspired approaches and validation of k-nearest neighbor classification results.

The results for Case 1 identified 276, 321, 324, 302 and 265 outlier domains using FA, Chaotic FA (Tuning μ), Chaotic FA (Tuning A), BA and CS respectively. After manually examining the web pages a total of 302 web pages were actually found to be spam web pages from the training 1009 web pages. Spam websites had factors like high level of spun content, abnormal link structures in the pages, and high number of articles with plagiarized content on the home page which does not adhere to webmaster's generally accepted quality guidelines, as elaborated earlier.

The findings were resulting in an accuracy of 96.24%, 98.48%, 97.87%, 98.53% and 97.67% respectively for the approaches algorithms. The accuracy is highest for BA followed by Chaotic FA for tuning absorption coefficient (μ). However, when the convergence speeds of the approaches were analyzed, it was seen that the chaotic variants of FA converge fastest while the CS with levy flights takes the maximum number of iterations to converge. This re-asserts our underlying reason for considering chaotic maps in the proposed approach. The outlier plots along with convergence functions for Case-1 (Business Fundas) are illustrated in Figure 6.7 and Figure 6.8 respectively.

For the revalidation of the findings for Case 1, another round of assessment was conducted with Case 2. The chaotic firefly variants are known to converge faster than the original firefly variants when used to mine outliers for the Case-2 for Tech Talk as well. The convergence plots for the same are shown in Figure 6.9. It is seen that the chaotic versions of FA converge faster than the original firefly and show promising results in terms of accurately identifying the outliers.

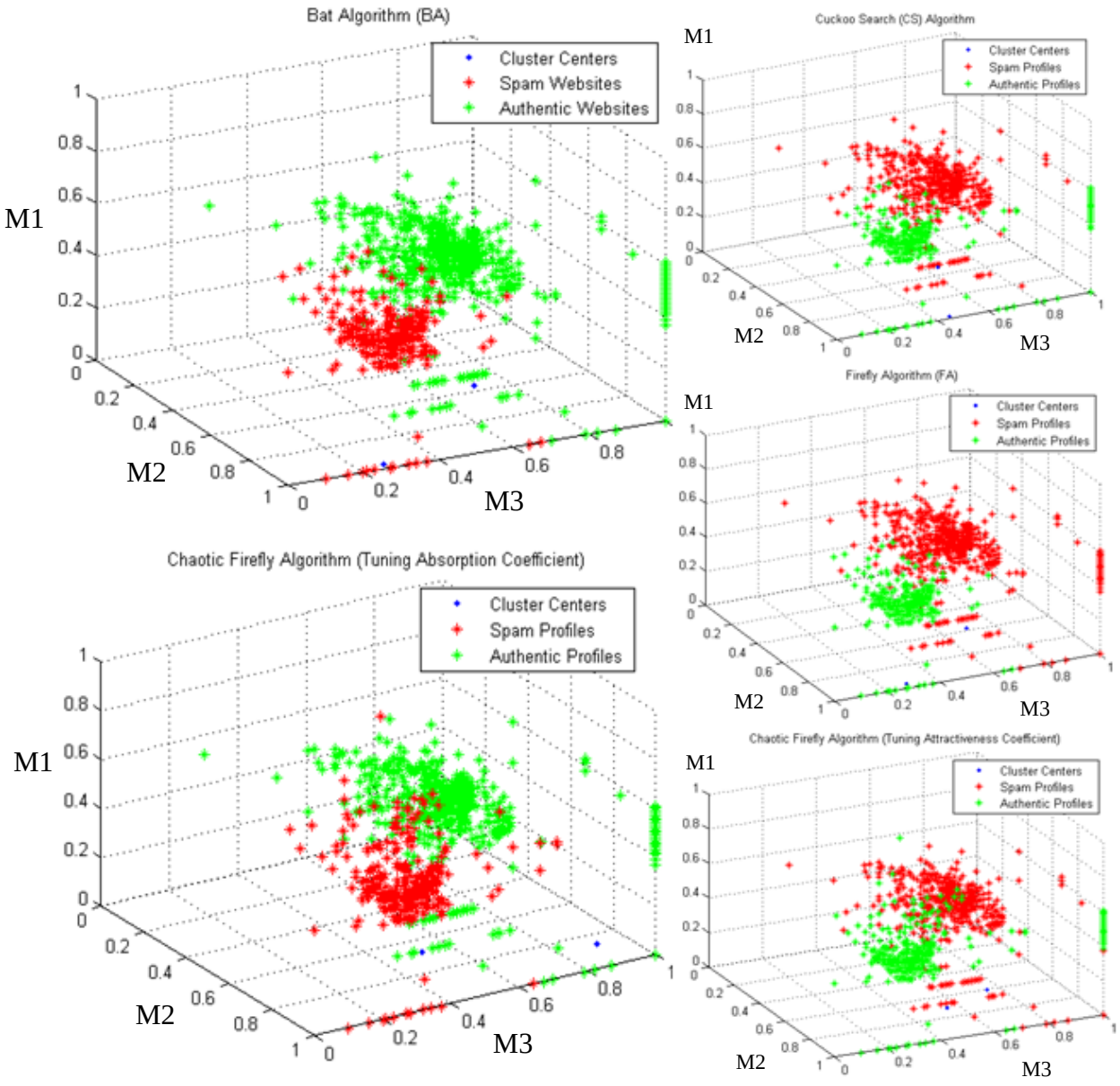


Figure 6.7: Outlier Plots for Case-I for Website Spam

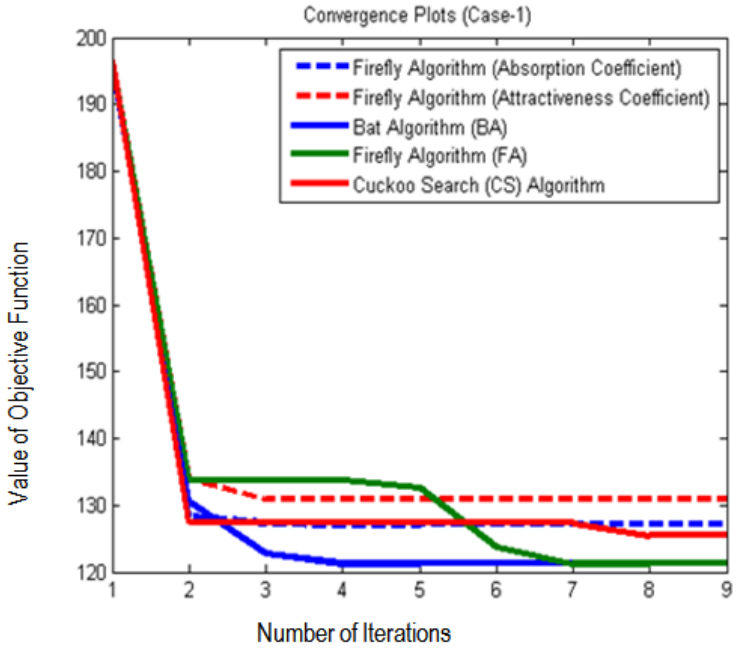


Figure 6.8: Comparative Convergence Plots for Case-1

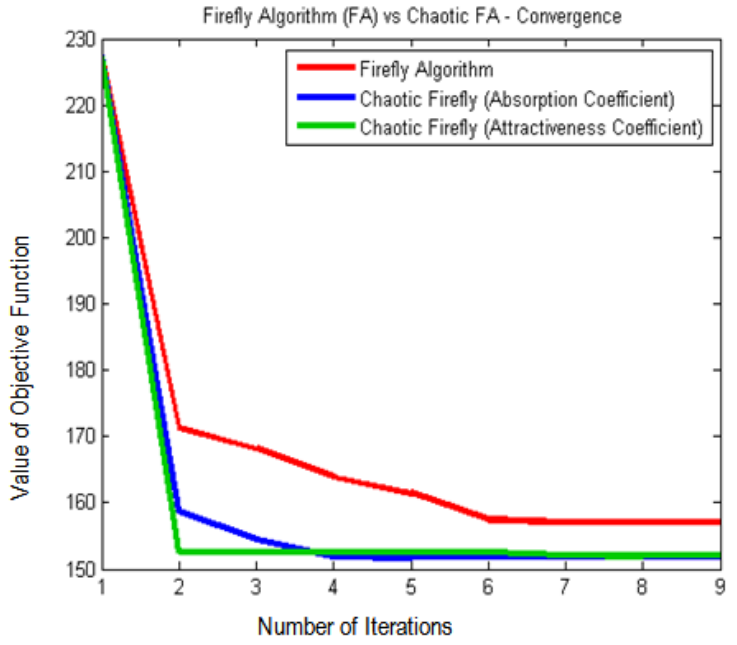


Figure 6.9: Convergence Plots for FA and Chaotic FA (Case-2)

For Case 2, the outlier profiles detected by FA, BA and CS were 147, 133 and 180 respectively. The actual spam profiles were 158 out of 642 training websites. This resulted into an accuracy of 97.61%, 96.39% and 95.71% for bat, firefly and cuckoo search algorithms respectively. The chaotic variants of firefly algorithm for tuning absorption and attractiveness coefficient give an accuracy of 98.23% and 97.34% respectively. The results for chaotic firefly algorithm for tuning the absorption coefficient for Case 2 along with convergence function are illustrated in Figure 6.10.

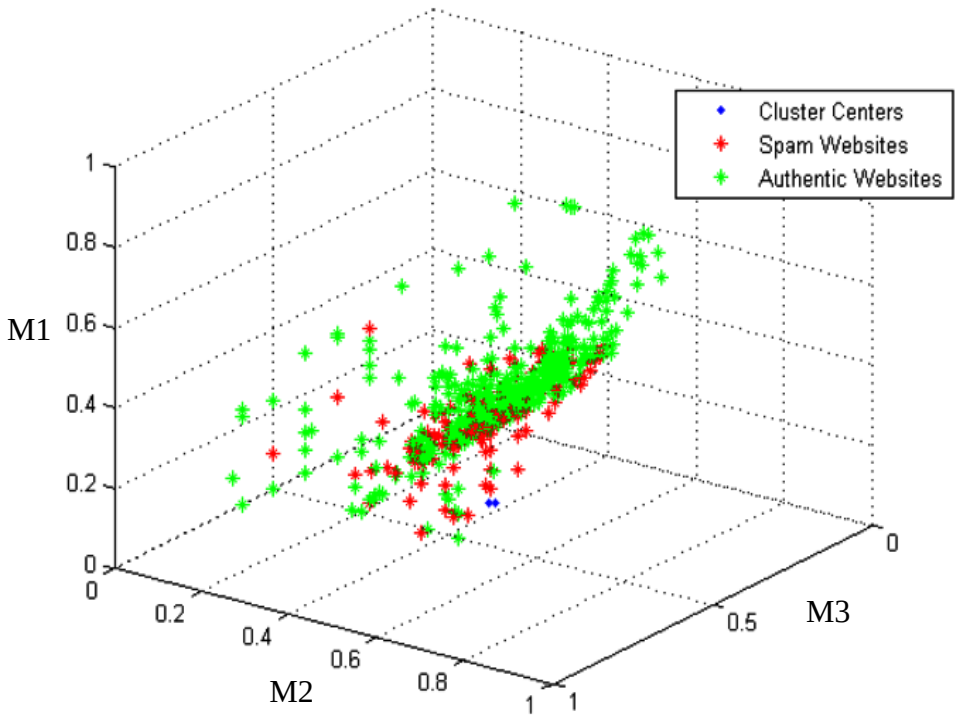


Figure 6.10: Outlier plots for Chaotic FA by Tuning Absorption Coefficient (Case-2)

Thus, it is evident that Firefly Algorithm (FA) gives a higher accuracy for both Case 1 of Business Fundas and Case 2 of Tech Talk. The dataset domain and data distribution play an important role in the performance on an approach. Here, since the datasets belong to similar domains and selection criteria, we have bat algorithm performing better than the rest of the two. Usually, when it comes to optimization algorithms, we cannot generalize on which approach is better. The No Free Lunch (NFL) theorem clearly states that an algorithm can perform

extremely well on a dataset and the same algorithm may give abysmal results on the other one [147]. This is because of varying data distributions and domains. Further, the k-nearest neighbor approach gave an accuracy of 96.59% for Case-1 and 97.47% for Case-2 using a manual evaluation of the websites for spam. The comparative results for Case I are illustrated in Table 6.6 with a comparison in terms of accuracy measures and convergence iterations.

Table 6.6: Comparative Results for Website Spam Identification

Approach	**Algorithm**	**Accuracy**	**Precision**	**Recall**	**F-measure**	**Convergence Iterations**
Proposed	Chaotic FA (Tuning Absorption)	**0.98**	0.94	**0.99**	**0.96**	**3**
	Chaotic FA (Tuning Attractiveness)	0.97	0.94	0.95	0.94	**3**
Existing	Firefly Algorithm (FA)	0.96	0.93	0.94	0.93	7
	Cuckoo Search (CS)	0.97	0.95	0.96	0.95	8
	Bat Algorithm (BA)	**0.98**	**0.95**	0.97	0.95	4

6.2.5 Conclusion and Future Research Scope

In the current scenario where there is plethora of information available on the web and people are in a constant quest to know more, the information visibility becomes of critical importance. Search engines are the greatest source for organic search. Websites are in a constant race to be on the top of search engine results. This makes assessment of websites for SEO critical since black hat SEO techniques including content spinning and link purchases are adopted by many

websites. This study attempts to identify spam websites that are used for link building in popular databases. The study uses two case studies and metrics for the same are identified. Case 1 for Business Fundas comprises of Page Rank (M1), Domain Authority (M2), Alexa Rank (M3), Social shares (M4) and Domain Age (M5) while Case 2 for Tech Talk includes Page Rank (M1), Page Authority (M2), Domain Authority (M3), Alexa Rank (M4), Google Index (M5), Social shares (M6) and Domain Age (M7). Further, methodologically, K-Means integrated chaotic firefly algorithm has been proposed to further detect website outliers. The proposed hybrid meta-heuristic approach is compared with K-Means integrated bat algorithm and cuckoo search algorithm for accuracy and computation speed. The K-means integrated bio inspired computing approaches used for mining outliers also are known to avoid locally optimum solutions that are common with traditional approaches discussed in literature [9] [38].

In addition to the methodological contribution, the study also collates several metrics based on off-site analytics for website evaluation in terms of factors like Page Rank, Page Authority, Domain Authority, MozRank, MozTrust, Citation Flow, Trust Flow, Alexa Rank, URL Rating, Domain Rating, Ahref Rank, Back Links, Total Links, Google Index, Social shares and Domain Age from various SEO companies like MajesticSEO, Moz, Ahref and webmaster tools. Metrics may be selected depending on the application and the knowledge portal under consideration for evaluating websites for any selection/rejection criteria. Thus, this study would be beneficial in the domain of vendor selection for SEO services to avoid investment on untrustworthy web sources for search engine marketing.

CHAPTER 7

7. CONCLUSION & FUTURE RESEARCH DIRECTIONS

In the current scenario where there is plethora of information available on the web and people are in a constant quest to know more, the information visibility becomes of critical importance. The Web 3.0 thus plays a key role in providing the content on the go. The work primarily targets social media and web analytics data to mine outliers. With the huge data influx, there are studies for outlier detection in high dimensional data. However, these approaches are computationally intensive often NP hard and also lead to a locally optimum solution. Since the data under consideration is huge and may also be unstructured textual data. This is creates need of integrating approaches that do not converge to a local optima. The meta-heuristic approaches are known to help in reaching to a globally optimum system.

The use of meta-heuristics specifically bio inspired computing techniques is also a novel contribution of this work. Further, bio inspired algorithms have been one of the most popular optimization techniques and mimic swarm behavior for optimization. Methodologically, the work integrates traditional k-Means and k-nearest neighbors approaches with bio inspired algorithms including firefly, cuckoo search, bat, grey wolf optimizer, artificial bee colony and wolf search algorithms for detecting outliers. The hybrid approaches are known to avoid local optimum. The work also uses chaos theory and Levy flight for better search space and faster results.

The work introduces chaos theory and integrates the same for detection of outliers. The chaotic algorithms use chaotic variables for random-based optimizations and are known to perform the overall search for an optimal solution at greater speeds. The primary reason for the faster speed is the non-repetition of chaos which avoida getting stuck into a single solution for a long time [27]. Also every meta-heuristic approach having stochastic components attempts to achieve by introducing some or the other probability distribution, common choices include uniform or

Gaussian distributions. However, instead of using these distributions it can be by principle advantageous to replace the same with chaotic maps. This is often a good choice since chaos possesses similar properties of randomness along with better statistical and dynamical properties. Such mixing properties ensure that the solutions generated by the algorithms are diverse enough to potentially reach every corner/mode in a multimodal landscape. Due to these dynamical properties of chaos, algorithms are potentially able to perform iterative search steps at higher speeds when compared to standard stochastic search methods having standard probability distributions.

Further, Levy Flights are also used in the work to enhance the accuracy of identification of outliers. Levy Flights have shown evidences of maximization of resource searches in uncertain environments. They are known to have infinite mean and variance and thus utilization of the same in meta-heuristic approaches greatly affects the accuracy of results. Meta-heuristic algorithms that integrate Levy Flight for reaching to the next potential solution can explore the search space more efficiently than algorithms that use standard Gaussian process. This advantage, combined search capabilities guarantees a global convergence with increased accuracy.

The work proposes several hybrid bio-inspired algorithms for detection of outliers. Although, the algorithms target selective datasets from the Web 3.0 domain, these can also be used in other domains with prior training of instances. However, when it comes to identifying the best optimization algorithm, it is not possible to generalize on which approach is better. The No Free Lunch (NFL) theorem for optimization techniques clearly states that an algorithm can perform extremely well on a dataset and the same algorithm may give abysmal results on the other one. The dataset domain and data distribution play an important role in the performance on an approach. The data set used for this analysis is large, but structured, and thus complexity driver is mostly in terms of time and number of iterations.

Future studies may integrate big data platforms to further reduce time complexity as well as handle unstructured data [176]. Other bio-inspired algorithms like monkey, lion, frog and wolf

may be integrated with traditional machine learning approaches to explore better results for different data distributions. In addition to that, since the results show significant overlap between the two clusters of spam and authentic marketing websites, a fuzzy based model may be used [178]. This would be beneficial in giving insights about the websites that have equal probability of being in the two clusters leading to a better classification accuracy than crisp clusters produced using the proposed approach. Future work is planned to use fuzzy clustering approaches along with the proposed approaches to highlight the overlapping clusters. The future work may also consider integrating the proposed approaches with big data platforms by using the Map Reduce Framework to further reduce the time required for executing the approaches.

REFERENCES

[1]. F. Garrigos, "Interrelationships between professional virtual communities and social networks, and the importance of virtual communities in creating and sharing knowledge," in *Connectivity and Knowledge Management in Virtual Organizations: Networking and Developing Interactive Communications*, 2009, IGI Global, pp. 1-22).

[2]. F. J. Garrigos-Simon, R. Lapiedra Alcami, & T. Barbera Ribera, "Social networks and Web 3.0: their impact on the management and marketing of organizations," *Management Decision*, vol. 50, no. 10, pp. 1880-1890, 2012.

[3]. J. Hendler, "Web 3.0 Emerging," *Computer*, vol. 42, no. 1, 2009.

[4]. Y. Zhang, N. Meratnia, & P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 2, pp. 159-170, 2010.

[5]. G. B. Gebremeskel, C. Yi, Z. He, & D. Haile, "Combined data mining techniques based patient data outlier detection for healthcare safety," *International Journal of Intelligent Computing and Cybernetics*, vol. 9, no. 1, pp. 42-68, 2016.

[6]. R. J. Bolton & D. J. Hand, "Unsupervised profiling methods for fraud detection," *Credit Scoring and Credit Control VII*, pp. 235-255, 2001.

[7]. V. Hodge & J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85-126, 2004.

[8]. A. Zimek, E. Schubert & H. P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," *Statistical Analysis and Data Mining*, vol. 5, no. 5, pp. 363-387, 2012.

[9]. V. Chandola, A. Banerjee & V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, pp. 41, vol. 3, pp. 15, 2009.

[10]. J. A. Hartigan & M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100-108, 1979.

[11]. T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman & A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE*

transactions on pattern analysis and machine intelligence, vol. 24, no. 7, pp. 881-892, 2002.

[12]. I. H. Osman & J. P. Kelly, "Meta-heuristics: an overview," in *Meta-heuristics*, Springer US, 1996, pp. 1-21.

[13]. D. F. Jones, S. K. Mirrazavi & M. Tamiz, "Multi-objective meta-heuristics: An overview of the current state-of-the-art," *European journal of operational research*, vol. 137, no. 1, pp. 1-9, 2002.

[14]. A. K. Kar, "Bio inspired computing–A review of algorithms and scope of applications," *Expert Systems with Applications*, vol. 59, pp. 20-32, 2016.

[15]. S. Grossberg, "Nonlinear neural networks: Principles, mechanisms, and architectures," Neural networks, vol. 1, no. 1, pp. 17–61, 1988.

[16]. J. H. Holland, "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence," in University of Michigan Press, 1975.

[17]. J. A. Snyman, "A new and dynamic method for unconstrained minimization," *Applied Mathematical Modelling*, vol. 6, no. 6, pp. 449–462, 1982.

[18]. M. M. Eusuff, K. E. Lansey, F. Pasha, "Shuffled frog-leaping algorithm: a memetic meta-heuristic for discrete optimization," Engineering Optimization, vol. 38, no. 2, pp. 129-154, 2006.

[19]. M. Dorigo, T. Stützle, "Ant Colony Optimization", Bradford Co., Scituate, MA, USA, 2004.

[20]. J. Kennedy, R. C. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Network*, pp. 1942–1948, 1995.

[21]. S. Das, A. Biswas, S. Dasgupta, A. Abraham, "Bacterial foraging optimization algorithm. Theoretical foundations, analysis, and applications," in *Foundations of computational intelligence*, Berlin Heidelberg, Springer, pp. 23–55, 2009.

[22]. X. S. Yang & S. Deb, "Cuckoo search via Lévy flights," in *World Congress on Nature & Biologically Inspired Computing, 2009. NaBIC 2009.* pp. 210-214).

[23]. D. Karaboga & B. Basturk, "A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm," *Journal of global optimization*, vol. 39, no. 3, pp. 459-471, 2009.

[24]. X. S. Yang, "Firefly algorithm, stochastic test functions and design optimization," *International Journal of Bio-Inspired Computation*, vol. 2, no. 2, pp. 78-84, 2010.

[25]. X. S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature inspired cooperative strategies for optimization*, Berlin: Springer, pp. 65–74, 2010.

[26]. S. Mirjalili, S. M. Mirjalili & A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46-61, 2014.

[27]. L. dos Santos Coelho & V. C. Mariani, "Use of chaotic sequences in a biologically inspired algorithm for engineering design optimization," *Expert Systems with Applications*, vol. 34, no. 3, pp. 1905-1913, 2008.

[28]. A. Cuzzocrea, I. Y. Song & K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!," in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*, October 2011, pp. 101-104.

[29]. D. M. Hawkins, "*Identification of outliers* (Vol. 11)," London: Chapman and Hall, 1980.

[30]. Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection* (Vol. 589). John wiley & sons.

[31]. V. Barnett & T. Lewis, *Outliers in statistical data*. Wiley, 1974.

[32]. A. S. Hadi, "A modification of a method for the detection of outliers in multivariate samples," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 393-396, 1994.

[33]. S. Papadimitriou, H. Kitagawa, P. B. Gibbons & C. Faloutsos, (2003, March). Loci: Fast outlier detection using the local correlation integral," in *19th International Conference on Data Engineering, March 2003*, pp. 315-326.

[34]. L. Kaufman & P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, Vol. 344, John Wiley & Sons, 2009.

[35]. R. T. Ng & J. Han, "CLARANS: A method for clustering objects for spatial data mining," *IEEE transactions on knowledge and data engineering*, vol. 14, no. 5, pp. 1003-1016, 2002.

[36]. D. Barbará & P. Chen, "Using the fractal dimension to cluster datasets," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, August, 2000*, pp. 260-264.

[37]. S. Papadopoulos, Y. Kompatsiaris, A. Vakali & P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515-554, 2012.

[38]. X. S. Yang, "A new metaheuristic bat-inspired algorithm," *Nature inspired cooperative strategies for optimization (NICSO 2010)*, Springer Berlin Heidelberg, pp. 65-74.

[39]. X. S. Yang, "Firefly algorithm, stochastic test functions and design optimization," *International Journal of Bio-Inspired Computation*, vol. 2, no. 2, 78-84, 2010.

[40]. A. Patcha & J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, vol. 51, no. 12, pp. 3448-3470, 2007.

[41]. J. Park, J. Kim & J. H. Lee, "Keyword extraction for blogs based on content richness," *Journal of Information Science*, vol. 40, 1, pp. 38-49, 2014.

[42]. I. H. Osman & J. P. Kelly, "Meta-heuristics: an overview," in *Meta-heuristics*, Springer US, 1996, pp. 1-21.

[43]. P. Grover & A. K. Kar, "Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature," *Global Journal of Flexible Systems Management*, vol. 18, no. 3, pp. 203–229, 2017.

[44]. A. Sahoo & S. Chandra, "Meta-heuristic approaches for active contour model based medical image segmentation," *International Journal of Advances in Soft Computing and Its Applications*, vol. 6, no. 2, 2014.

[45]. H. A. Bouarara, R. M. Hamou, A. Rahmani & A. Amine, "Application of Meta-Heuristics Methods on PIR Protocols Over Cloud Storage Services," *International Journal of Cloud Applications and Computing (IJCAC)*, vol. 4, no. 3, pp. 1-19, 2014.

[46]. M. Mavrovouniotis, C. Li & S. Yang, "A survey of swarm intelligence for dynamic optimization: algorithms and applications," *Swarm and Evolutionary Computation*, vol. 33, pp. 1-17, 2017.

[47]. A. Chakraborty & A. K. Kar, A. K, "A Review of Bio-Inspired Computing Methods and Potential Applications," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems, 2016*, Springer India, pp. 155-161.

[48]. A. Chakraborty & A. K. Kar, A. K, "Swarm intelligence: A review of algorithms," in *Nature-Inspired Computing and Optimization, 2017*, Springer International Publishing, pp. 475-494.

[49]. S. J. Nanda & G. Panda, "A survey on nature inspired metaheuristic algorithms for partitional clustering," *Swarm and Evolutionary computation*, vol. 16, pp. 1-18, 2014.

[50]. D. Binu, "Cluster analysis using optimization algorithms with newly designed objective functions," *Expert Systems with Applications*, vol. 42, no. 14, pp. 5848-5859, 2015.

[51]. R. Tang, S. Fong, X. S. Yang & S. Deb, "Wolf search algorithm with ephemeral memory," in *Seventh International Conference on Digital Information Management (ICDIM), Macao, 2012*, pp. 165-172.

[52]. M. Yazdani & F. Jolai, "Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm," *Journal of computational design and engineering*, vol. 3, no. 1, pp. 24-36, 2016.

[53]. X. Zhang, S. Huang, Y. Hu, Y. Zhang, S. Mahadevan & Y. Deng, "Solving 0-1 knapsack problems based on amoeboid organism algorithm," *Applied Mathematics and Computation*, vol. 219, no. 19, pp. 9959-9970, 2013.

[54]. X. S. Yang, "Flower pollination algorithm for global optimization," in *International conference on unconventional computing and natural computation, September, 2012*, Springer, Berlin, Heidelberg, pp. 240-249.

[55]. M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalhaim & S. Ur, "The shark-search algorithm. An application: tailored Web site mapping," *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 317-326, 1998.

[56]. G. Theraulaz, "Task differentiation in Polistes wasp colonies: a model for self-organizing groups of robots," in *Proceedings of the First International Conference on Simulation of Adaptive Behavior: From Animals to Animates, 1991*, The MIT Press, pp. 346-355.

[57]. M. L. Shahreza, D. Moazzami, B. Moshiri & M. R. Delavar, "Anomaly detection using a self-organizing map and particle swarm optimization," *Scientia Iranica*, vol. 18, no. 6, pp. 1460-1468, 2011.

[58]. M. H. A. Adaniya, T. Abrao & M. L. Proenca Jr, "Anomaly detection using metaheuristic firefly harmonic clustering," *Journal of Networks*, vol. 8, no. 1, pp. 82-91, 2013.

[59]. L. F. Carvalho, J. J. Rodrigues, S. Barbon & M. L. Proenca, "Using ant colony optimization metaheuristic and dynamic time warping for anomaly detection," in *21st International Conference on Software, Telecommunications and Computer Networks (SoftCOM), September 2013*, pp. 1-5.

[60]. G. Wang, J. Hao, J. Ma & L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert systems with applications*, vol. 37, no. 9, pp. 6225-6232, 2010.

[61]. T. F. Ghanem, W. S. Elkilani & H. M. Abdul-Kader, "A hybrid approach for efficient anomaly detection using metaheuristic methods," *Journal of advanced research*, vol. 6, no. 4, pp. 609-619, 2015.

[62]. P. DiMaggio & E. Hargittai, "From the 'digital divide'to 'digital inequality': Studying Internet use as penetration increases," *Princeton: Center for Arts and Cultural Policy Studies, Woodrow Wilson School, Princeton University*, vol. 4, no. 1, pp. 4-2, 2001.

[63]. A. J. Flanagin & M. J. Metzger, "Internet use in the contemporary media environment," *Human communication research*, vol. 27, no. 1, pp. 153-181, 2001.

[64]. V. Shah, R. Nojin Kwak, D. Lance Holbert, “"Connecting" and "disconnecting" with civic life: Patterns of Internet use and the production of social capital,” *Political communication*, vol. 18, no. 2, pp. 141-162, 2001.

[65]. Z. Papacharissi & A. M. Rubin, “Predictors of Internet use,” *Journal of broadcasting & electronic media*, vol. 44, no. 2, pp. 175-196, 2000.

[66]. F. Wu, V. Mahajan & S. Balasubramanian, “An analysis of e-business adoption and its impact on business performance,” *Journal of the Academy of Marketing science*, vol. 31, no. 4, pp. 425-447, 2003.

[67]. G. J. Simmons, “i-Branding: developing the internet as a branding tool,” *Marketing Intelligence & Planning*, vol. 25, no. 6, pp. 544-562, 2007.

[68]. T. C. Powell & A. Dent-Micallef, “Information technology as competitive advantage: The role of human, business, and technology resources,” *Strategic management journal*, vol. 18, no. 5, pp. 375-405, 1997.

[69]. R. Amit, and C. Zott, C., “Value creation in e-business,” *Strategic management journal*, vol. 22, no. 6-7, pp.493-520, 2001.

[70]. R. A. Malaga, “Web 2.0 Techniques for search engine optimization: Two case studies,” *Review of Business Research*, vol. 9, no. 1, pp. 132-139, 2009.

[71]. A. S. Abrahams, J. Jiao, W. Fan, G. A. Wang & Z. Zhang, “What's buzzing in the blizzard of buzz? Automotive component isolation in social media postings,” *Decision Support Systems*, vol. 55, no. 4, pp. 871-882, 2013.

[72]. J. McCarthy, J. Rowley, C. Jane Ashworth & E. Pioch, “Managing brand presence through social media: the case of UK football clubs,” *Internet Research*, vol. 24, no. 2, pp. 181-204, 2014.

[73]. J. H. Kietzmann, K. Hermkens, I. P. McCarthy & B. S. Silvestre, “Social media? Get serious! Understanding the functional building blocks of social media,” *Business horizons*, vol. 54, no. 3, pp. 241-251, 2011.

[74]. A. Lenhart, K. Purcell, A. Smith & K. Zickuhr, “Social Media & Mobile Internet Use among Teens and Young Adults,” Millennials. *Pew internet & American life project*, 2010.

[75]. W. G. Mangold & D. J. Faulds, "Social media: The new hybrid element of the promotion mix," *Business horizons*, vol. 52, no. 4, pp. 357-365, 2009.

[76]. S. Aral & D. Walker, "Creating social contagion through viral product design: A randomized trial of peer influence in networks," *Management science*, vol. 57, no. 9, pp. 1623-1639.

[77]. H. Gil de Zúñiga, N. Jung & S. Valenzuela, "Social media use for news and individuals' social capital, civic engagement and political participation," *Journal of Computer-Mediated Communication*, vol. 17, no. 3, pp. 319-336, 2012.

[78]. G. Khatwani, O. Anand & A. K. Kar, "Evaluating internet information search channels using hybrid MCDM technique," in *International Conference on Swarm, Evolutionary, and Memetic Computing, December 2014*, Springer International Publishing, pp. 123-133.

[79]. R. Hanna, A. Rohm & V. L. Crittenden, "We're all connected: The power of the social media ecosystem," *Business horizons*, vol. 54, no. 3, pp. 265-273, 2011.

[80]. A. H. Zadeh & R. Sharda, "Modeling brand post popularity dynamics in online social networks," *Decision Support Systems*, vol. 65, pp. 59-68, 2014.

[81]. C. Shirky, "The political power of social media: Technology, the public sphere, and political change," *Foreign affairs*, pp. 28-41, 2011.

[82]. A. Patino, D. A. Pitta & R. Quinones, "Social media's emerging importance in market research," *Journal of Consumer Marketing*, vol. 29, no. 3, pp. 233-237, 2012.

[83]. X. Han, L. Wang, N. Crespi, S. Park & A. Cuevas, "Alike people, alike interests? Inferring interest similarity in online social networks," *Decision Support Systems*, vol. 69, pp. 92-106, 2015.

[84]. T. Nagle & A. Pope, A. (2013), "Understanding social media business value, a prerequisite for social media selection," *Journal of Decision Systems*, vol. 22, no. 4, pp. 283-297, 2013.

[85]. M. Y. Kiang, T. S. Raghu & K. H. M. Shang, "Marketing on the Internet—who can benefit from an online marketing approach?," *Decision Support Systems*, vol. 27, no. 4, pp. 383-393, 2000.

[86]. S. Brin & L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825-3833, 2012.

[87]. C. W. Choo, B. Detlor & D. Turnbull, "Information seeking on the Web: An integrated model of browsing and searching," *first monday*, vol. 5, no. 2, 2000.

[88]. B. J. Jansen & A. Spink, "How are we searching the World Wide Web? A comparison of nine search engine transaction logs," *Information processing & management*, vol. 42, no. 1, pp. 248-263, 2006.

[89]. R. D. Young, Who Uses Search Engines? 92% of Adult U.S. Internet Users. Search Engine Watch. https://searchenginewatch.com/sew/study/2101282/search-engines-92-adult-internet-users-study, accessed 15 February, 2017.

[90]. R. Sen, "Optimal search engine marketing strategy," *International Journal of Electronic Commerce*, vol. 10, no. 1, pp. 9-25, 2005.

[91]. H. Catherine Murphy & C. D. Kielgast, "Do small and medium-sized hotels exploit search engine marketing?," *International Journal of Contemporary Hospitality Management*, vol. 20, no. 1, pp. 90-97, 2008.

[92]. B. J. Jansen & P. R. Molina, "The effectiveness of Web search engines for retrieving relevant ecommerce links," *Information Processing & Management*, vol. 42, no. 4, pp. 1075-1098, 2006.

[93]. B. Pan, Z. Xiang, R. Law, & D. R Fesenmaier, "The dynamics of search engine marketing for tourist destinations," *Journal of Travel Research*, vol. 50, no. 4, pp. 365-377, 2011.

[94]. W. Dou, K. H. Lim, C. Su, N. Zhou & N. Cui, "Brand positioning strategy using search engine marketing," *Mis Quarterly*, pp. 261-279, 2010.

[95]. G. Spais, ""Search Engine Optimization (SEO) as a dynamic online promotion technique: The implications of activity theory for promotion managers," *Innovative Marketing*, vol. 6, no. 1, pp. 7-24, 2010.

[96]. J. B. Killoran, "How to use search engine optimization techniques to increase website visibility," *IEEE Transactions on professional communication*, vol. 56, no. 1, 50-66, 2013.

[97]. R. A. Malaga, "Search engine optimization—black and white hat approaches," *Advances in Computers*, vol. 78, pp. 1-39, 2010.

[98]. R. A. Malaga, "Worst practices in search engine optimization," *Communications of the ACM*, vol. 51, no. 12, pp. 147-150, 2008.

[99]. E. Acuna & C. Rodriguez, "A meta-analysis study of outlier detection methods in classification," *Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez*, 2004.

[100]. S. Alam, G. Dobbie, P. Riddle & M. A. Naeem, "A swarm intelligence based clustering approach for outlier detection," in *IEEE Congress on Evolutionary Computation (CEC), July 2010*, pp. 1-7.

[101]. A. Forestiero, "Bio-inspired algorithm for outliers detection," *Multimedia Tools and Applications*, pp. 1-19, 2017.

[102]. P. Murugavel & M. Punithavalli, "Improved hybrid clustering and distance-based technique for outlier removal," *International Journal on Computer Science and Engineering (IJCSE)*, vol. 3, no. 1, pp. 333-339, 2011.

[103]. P. S. Leeflang, P. C. Verhoef, P. Dahlström, P., & Freundt, T. (2014). Challenges and solutions for marketing in a digital era. *European management journal*, *32*(1), 1-12.

[104]. W. Dou, K. H. Lim, C. Su, N. Zhou, & N. Cui, "Brand positioning strategy using search engine marketing," *Mis Quarterly*, pp. 261-279, 2010.

[105]. M. Sawhney, G. Verona & E. Prandelli, "Collaborating to create: The Internet as a platform for customer engagement in product innovation," *Journal of interactive marketing*, vol. 19, no. 4, pp. 4-17, 2005.

[106]. N. Hayes, *Influencer Marketing: Who Really Influences Your Customers*. Taylor & Francis, 2008.

[107]. D. Chaffey & M. Patron, "From web analytics to digital marketing optimization: Increasing the commercial value of digital analytics," *Journal of Direct, Data and Digital Marketing Practice*, vol. 14, no. 1, pp. 30-45, 2012.

[108]. L. Moreno & P. Martinez, "Overlapping factors in search engine optimization and web accessibility," *Online Information Review*, vol. 37, no. 4, 564-580, 2013.

[109]. J. A. Slegg, "Complete Guide to Panda, Penguin, and Hummingbird. Search Engine Journal,"http://www.searchenginejournal.com/seo-guide/google-penguin-panda-hummingbird, last accessed 2017/02/15.

[110]. A. Jain & M. Dave, "The role of backlinks in search engine ranking," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 4, 2013.

[111]. H. Zuze & M. Weideman, "Keyword stuffing and the big three search engines," *Online Information Review*, vol. 37, no. 2, pp. 268-286, 2013.

[112]. Y. Lee & K. A. Kozar, "Investigating the effect of website quality on e-business success: An analytic hierarchy process (AHP) approach," *Decision support systems*, vol. 42, no. 3, pp. 1383-1401, 2006.

[113]. M. P. Evans, "Analysing Google rankings through search engine optimization data," *Internet research*, vol. 17, no. 1, 21-37, 2007.

[114]. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*," vol. 46, no. 5, 604-632, 1999.

[115]. L. Page, S. Brin, R. Motwani & T. Winograd, *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.

[116]. Z. Gyöngyi, H. Garcia-Molina & J. Pedersen, "Combating web spam with trustrank," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, August 2004*, VLDB Endowment, pp. 576-587.

[117]. A. K. Kar, "A Decision Support System for Website Selection for Internet Based Advertising and Promotions," in: Sengupta S., Das K., Khan G. (eds), in *Emerging Trends in Computing and Communication*, Lecture Notes in Electrical Engineering, vol. 298, Springer, 2014.

[118]. S. Binitha & S. S. Sathya, "A survey of bio inspired optimization algorithms," *International Journal of Soft Computing and Engineering*, vol. 2, no. 2, pp. 137-151, 2012.

[119]. R. Tang, S. Fong, X. S. Yang & S. Deb, "Integrating nature-inspired optimization algorithms to K-means clustering," in *2012 Seventh International Conference on Digital Information Management (ICDIM), August, 2012*, pp. 116-123.

[120]. T. Utsuro, C. Zhao, L. Xu, J. Li & Y. Kawada, "An Empirical Analysis on Comparing Market Share with Concerns on Companies Measured Through Search Engine Suggests," *Global Journal of Flexible Systems Management*, pp. 1-17, 2017.

[121]. I. Papasolomou & Y. Melanthiou, "Social media: Marketing public relations' new best friend," *Journal of Promotion Management*, vol. 18, no. 3, pp. 319-328, 2012.

[122]. Y. Li, M. Qian, D. Jin, P. Hui & A. V. Vasilakos, "Revealing the efficiency of information diffusion in online social networks of microblog," *Information Sciences*, vol. 293, pp. 383-389, 2015.

[123]. G. S. O'Keeffe & K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800-804, 2011.

[124]. T. Fisher, "ROI in social media: A look at the arguments," *The Journal of Database Marketing & Customer Strategy Management*, vol. 16, no. 3, pp. 189-195, 2009.

[125]. D. L. Hoffman & M. Fodor, "Can you measure the ROI of your social media marketing?," *MIT Sloan Management Review*, vol. 52, no. 1, pp. 41, 2010.

[126]. V. Kumar & R. Mirchandani, "Increasing the ROI of social media marketing," *MIT Sloan Management Review*, vol. 54, no. 1, 55, 2012.

[127]. S. Edosomwan, S. K. Prakasan, D. Kouame, J. Watson & T. Seymour, "The history of social media and its impact on business," *Journal of Applied Management and entrepreneurship*, vol. 16, no. 3, pp. 79, 2011.

[128]. A. M. Kaplan & M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business horizons*, vol. 53, no. 1, pp. 59-68, 2010.

[129]. B. K. Chae, "Insights from hashtag# supplychain and Twitter Analytics: Considering Twitter and Twitter data for supply chain practice and research," *International Journal of Production Economics*, vol. 165, pp. 247-259, 2015.

[130]. J. Berger & K. Milkman, "Virality: What Gets Shared and Why," *Advances in Consumer Research*, vol. 37, 2010.

[131]. M. Guerini, C. Strapparava & G. Özbal, "Exploring Text Virality in Social Networks," in *ICWSM, July*, 2011.

[132]. J. Colliander & M. Dahlén, "Following the fashionable friend: The power of social media," *Journal of advertising research*, vol. 51, no. 1, pp. 313-320, 2011.

[133]. X. Luo & J. Zhang, "How do consumer buzz and traffic in social media marketing predict the value of the firm?," *Journal of Management Information Systems*, vol. 30, no. 2, pp. 213-238, 2013.

[134]. P. R. Berthon, L. F. Pitt, K. Plangger & D. Shapiro, D. (2012). Marketing meets Web 2.0, social media, and creative consumers: Implications for international marketing strategy," *Business horizons*, vol. 55, no. 3, pp. 261-271, 2012.

[135]. A. Hausmann, "Creating 'buzz': opportunities and limitations of social media for arts institutions and their viral marketing," *International Journal of Nonprofit and Voluntary Sector Marketing*, vol. 17, no. 3, pp. 173-182, 2012.

[136]. J. Leskovec, L. A. Adamic & B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, pp. 5, 2007.

[137]. D. Centola, "The spread of behavior in an online social network experiment," *science*, vol. 329, no. 5996, pp. 1194-1197, 2010.

[138]. L. Weng, F. Menczer & Y. Y. Ahn, "Virality prediction and community structure in social networks," *Scientific reports*. *3*, 2013.

[139]. H. Kwak, C. Lee, H. Park & S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web, April 2010*, pp. 591-600, 2010.

[140]. A. M. Popescu & M. Pennacchiotti, "Detecting controversial events from twitter," in *Proceedings of the 19th ACM international conference on Information and knowledge management, October 2010*, pp. 1873-1876.

[141]. D. Murthy, "Twitter and elections: are tweets, predictive, reactive, or a form of buzz?," *Information, Communication & Society*, vol. 18, no. 7, pp. 816-831, 2015.

[142]. F. Thies, M. Wessel & A. Benlian, "Understanding the dynamic interplay of social buzz and contribution behavior within and between online platforms–evidence from crowdfunding, 2014.

[143]. X. Zhang, H. Fuehres & P. Gloor, "Predicting asset value through twitter buzz," *Advances in Collective Intelligence 2011*, pp. 23-34, 2012.

[144]. F. Kawala, A. Douzal-Chouakria, E. Gaussier & E. Dimert, "Prédictions d'activité dans les réseaux sociaux en ligne," in *4ième Conférence sur les Modèles et l'Analyse des Réseaux: Approches Mathématiques et Informatiques, October, 2013*, p. 16.

[145]. D. Karaboga & B. Akay, "A comparative study of artificial bee colony algorithm," *Applied mathematics and computation*, vol. 214, no. 1, pp. 108-132, 2009.

[146]. D. Karaboga & B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied soft computing*, vol. 8, no. 1, pp. 687-697, 2008.

[147]. D. H. Wolpert & W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67-82, 1997.

[148]. S. Gurajala, J. S. White, B. Hudson & J. N. Matthews, "Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach," in *Proceedings of the 2015 International Conference on Social Media & Society, July 2015*, pp. 9.

[149]. M. Fire, D. Kagan, A. Elyashar & Y. Elovici, "Friend or foe? Fake profile identification in online social networks," *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 194, 2014.

[150]. S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi & M. Tesconi, "Fame for sale: efficient detection of fake Twitter followers," *Decision Support Systems*, vol. 80, pp. 56-71, 2015.

[151]. N. Dalkey & O. Helmer, "An experimental application of the Delphi method to the use of experts," *Management science*, vol. 9, no. 3, pp. 458-467, 1963.

[152]. J. L. Aaker, "Dimensions of brand personality," *Journal of marketing research*, pp. 347-356, 1997.

[153]. A. Bruns & S. Stieglitz, "Towards more systematic Twitter analysis: Metrics for tweeting activities," *International Journal of Social Research Methodology*, vol. 16, no. 2, pp. 91-108, 2013.

[154]. A. Kishor & P. K. Singh, "Empirical study of grey wolf optimizer," in *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*. Springer, Singapore, 2016, pp. 1037-1049.

[155]. W. J. Conover & R. L. Iman, "Rank transformations as a bridge between parametric and nonparametric statistics," *The American Statistician*, vol. 35, no. 3, 124-129, 1981.

[156]. H. Levene, "Robust tests for equality of variances," *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, vol. 2, pp. 278-292, 1960.

[157]. P. Galán-García, J G. D. L. Puerta, C. L. Gómez, C. L., I. Santos & P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network:

Application to a real case of cyberbullying," *Logic Journal of the IGPL*, vol. 24, no. 1, pp. 42-53, 2016.

[158]. P. Cunningham & S. J. Delany, "k-Nearest neighbour classifiers," *Multiple Classifier Systems*, vol. 34, pp. 1-17, 2007.

[159]. J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun & J. Han, "On community outliers and their efficient detection in information networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, July 2010*, pp. 813-822.

[160]. P. T. Metaxas & E. Mustafaraj, "Social media and the elections," *Science*, vol. 338, no. 6106, pp. 472-473, 2012.

[161]. G. Brown, T. Howe, M. Ihbe, A. Prakash & K. Borders, "Social networks and context-aware spam," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work, Novmber, 2008*, pp. 403-412.

[162]. J. Y. Ho & M. Dempsey, "Viral marketing: Motivations to forward online content," *Journal of Business research*, vol. 63, no. 9, 1000-1006, 2010.

[163]. G. G. Wang, S. Deb, A. H. Gandomi, Z. Zhang & A. H. Alavi, "Chaotic cuckoo search," *Soft Computing*, vol. 20, no. 9, pp. 3349-3362, 2016.

[164]. G. Szabo & B. A. Huberman, "Predicting the popularity of online content," *Communications of the ACM*, vol. 53, no. 8, pp. 80-88, 2010.

[165]. M. Cha, H. Kwak, P. Rodriguez, Y. Y. Ahn & S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Transactions on Networking (TON)*, vol. 17, no. 5, pp. 1357-1370, 2009.

[166]. A. Tatar, J. Leguay, P. Antoniadis, A. Limbourg, M. D. de Amorim & S. Fdida, "Predicting the popularity of online articles based on user comments," in *Proceedings of the International Conference on Web Intelligence, Mining and Semantics, May 2011*, pp. 67.

[167]. J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer & A. Vespignani, "Characterizing and modeling the dynamics of online popularity," *Physical review letters*, vol. 105, no. 15, 2010

[168]. K. Fernandes, P. Vinagre & P. Cortez, "A proactive intelligent decision support system for predicting the popularity of online news," in *Portuguese Conference on Artificial Intelligence, September 2015*, Springer, Cham, pp. 535-546.

[169]. X. S. Yang & S. Deb, "Engineering optimisation by cuckoo search," *International Journal of Mathematical Modelling and Numerical Optimisation*, vol. 1, no. 4, pp. 330-343, 2010.

[170]. R. N. Mantegna & H. E. Stanley, "Stochastic process with ultraslow convergence to a Gaussian: the truncated Lévy flight," *Physical Review Letters*, vol. 73, no. 22, pp. 2946, 1994.

[171]. A. H. Gandomi, X. S. Yang & A. H. Alavi, "Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems," *Engineering with computers*, vol. 29, no. 1, pp. 17-35, 2013.

[172]. J. A. Malcolm & P. C. Lane, "An approach to detecting article spinning," in *Proceedings of the Third International Conference on Plagiarism*, 2008.

[173]. S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto ... & K. P. Gummadi, "Understanding and combating link farming in the twitter social network," in *Proceedings of the 21st international conference on World Wide Web, April 2012*, pp. 61-70.

[174]. A. K. Kar, "A Decision support system for website selection for internet based advertising and promotions," in *Emerging Trends in Computing and Communication*, Springer India, *2014*, pp. 453-457.

[175]. A. H. Gandomi, X. S. Yang, S. Talatahari & A. H. Alavi, "Firefly algorithm with chaos," *Communications in Nonlinear Science and Numerical Simulation*, vol. 18, no. 1, pp. 89-98, 2013.

[176]. L. Gu & H. Li, "Memory or time: Performance evaluation for iterative operation on hadoop and spark," in *IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC)* , *November 2013*, pp. 721-727.

[177]. P. Green, The content of a college-level outdoor leadership course, 1982.

[178]. J. C. Bezdek, R. Ehrlich & W. Full, “FCM: The fuzzy c-means clustering algorithm,” *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[179]. I. Fister, X. S. Yang & J. Brest, “A comprehensive review of firefly algorithms,” *Swarm and Evolutionary Computation*, vol. 13, pp. 34-46, 2013.

[180]. J. Senthilnath, S. N. Omkar & V. Mani, “Clustering using firefly algorithm: performance study,” *Swarm and Evolutionary Computation*, vol. 1, no. 3, pp. 164-171, 2011.

PUBLICATIONS

Journals published online:

1. Aswani, R., Ghrera, S. P., & Chandra, S. (2016). A Novel Approach to Outlier Detection using Modified Grey Wolf Optimization and k-Nearest Neighbors Algorithm. *Indian Journal of Science and Technology*, *9*(44). **[SCOPUS]**

2. Aswani, R., Ghrera, S. P., Kar, A. K., & Chandra, S. (2017). Identifying buzz in social media: a hybrid approach using artificial bee colony and k-nearest neighbors for outlier detection. *Social Network Analysis and Mining*, *7*(1), 38. **[SCOPUS, Emerging SCI]**

Conference/Book Chapters:

1. Aswani, R., Ghrera, S. P., Chandra, S., & Kar, A. K. (2017, November). Outlier Detection among Influencer Blogs Based on off-Site Web Analytics Data. In *Conference on e-Business, e-Services and e-Society* (pp. 251-260). Springer, Cham. **[SCOPUS]**

2. Aswani, R., Ghrera, S.P., Chandra, S., & Kar, A.K. (2017, December). Identifying popular online news: An approach using chaotic cuckoo search algorithm. Forthcoming in International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), IEEE **[SCOPUS]**

3. Further, the first published journal article "A Novel Approach to Outlier Detection using Modified Grey Wolf Optimization and k-Nearest Neighbors Algorithm" was also presented at SHANNON 100 -THIRD INTERNATIONAL CONFERENCE held on 8th-9th April 2016 at Lovely Professional University, Jalandhar.

Journals under review:

1. Aswani, R., Ghrera, S. P., Kar, A. K., & Chandra, S. (2018). A hybrid evolutionary approach for identifying spam websites for search engine marketing. *Evolutionary Intelligence* **[Revise & Resubmit]**

2. Aswani, R., Ghrera, S. P., Kar, A. K., & Chandra, S. (2018). Identifying Fake Profiles in Social Media – Insights from Twitter Analytics. *IEEE Transactions on Computational Social Systems* **[Revise & Resubmit]**

RESPONSE TO REVIEWERS

Comments from Reviewer-1:

Following are some suggestions to improve the quality of PhD Thesis:

1. **In Chapter 2 conclusion may be included that summarizes major research limitations of the existing approaches and that forms basis for the proposed research.**

   Response: I thank the reviewer for pointing out such a critical part that was missing. A section on Conclusion has been added at the end of Chapter 2, in the form of Section 2.4. The section summarizes the critical research gaps that are addressed in the subsequent methodology chapter using different case scenarios.

2. **In Section 4.1.4, it has been shown that the proposed approach behaves better than other existing approaches; however, detailed justification for better results is missing. So, justification be given, wherever required.**

   Response: I appreciate the reviewers comment for adding a detailed justification for the better performance of the proposed approach. The same has been added in a paragraph. This justification definitely enhances the readability and provides theoretical foundations as to why the approach has been used and is performing better.

3. **In Chapter 4, axis titles/labels in various graphs are missing.**

   Response: The titles/labels have been added for all the graphs.

4. **Tables are divided across the pages. Preferably, a table must fit within a page.**

   Response: The tables have been restricted to a single place wherever it was possible. However, some tables are exceeding the page by a lot of lines and thus cannot be fit within a page. For these tables it has been ensured that the rows do not split in different pages.

5. **References must follow standard IEEE format.**

Response: The references have been formatted in the desired IEEE format.

6. **There are many typographic errors that need to be corrected. There is requirement of substantial improvement in the Grammar.**
   Response: I have tried my best to rectify the typographic and grammatical errors by doing a thorough proof read of the thesis.

Comments from Reviewer-2:

The thesis presents a novel methodology for outlier detection, which is a hybrid of bio inspired computing algorithms with traditional machine learning algorithms namely, KNN and K-Means. The motivation for this research is to avoid locally optimum solutions and minimize the convergence time, both of which are major limitations of traditional machine learning algorithms, when used on large datasets. The thesis is divided appropriately into providing improvements in solving various related problems and sub-problems.

Positives:

• The proposed methodology for each task and the results obtained have a direct application in a range of industries which can bring a significant improvement in their existing systems.

• A detailed explanation is given for all the algorithms used. The findings of an approach each task are clearly mentioned & an appropriate analysis is provided for the same.

• The attributes and metrics for a dataset on each task is chosen after an in-depth study and a careful analysis of every aspect concerned.

• All the results are adequately represented in a tabular form wherever required, along with a comparative study of the proposed approach with the existing methods.

• Colorful graphs and diagrams enhance the interpretability of the results.

I thank the reviewer for highlighting the positives. This has truly motivated me to enhance my work further. I also thank the reviewer the constructive criticism expressed in the form of the points below. I have tried my best to address the concerns raised.

Possible Improvements:

1. **The scope of study can be expanded to include other machine learning algorithms like neural-networks which are a current trend in a hybrid with bio inspired algorithms. Or an explanation for not including them can be helpful.**

   Response: I understand the reviewers concern since neural networks are widely used in the literature for solving various problems. Neural networks do upgrade the potential solution iteratively using a feedback mechanism but they do not eliminate local optima since these are also traditional algorithms and can get stuck in the local solution. Further, to avoid this, some optimization approach will have to be integrated with them which will solve this problem. Tis however will increase the complexity of the solution by multifold. And it will take hours for the model to converge.

2. **For Outlier Detection in a Supervised Scenario, as the study is being carried with a focus on industries revolving around Web 3.0, a related dataset can be a better choice compared to the current biology based Iris and Abalone datasets.**

   Response: The Iris and Abalone datasets are standard datasets and have been used so that results could be validated. Further, this was done during the early phases of the work when I did not know the importance of self-collected data and the implications that I could provide in the domain as well. The proposed approach for the section has thus been slightly modified to be used in another case scenario surrounding fake profile identification which is more relevant for the domain at hand. I felt it would not be justified to remove this work from the thesis since it was already published and cited multiple times.

3. **The pseudo-codes can be less verbose and more symbolic. An alternative option can be to represent them as algorithms and not as pseudo-codes.**

Response: I understand the concerns of the reviewer in this aspect. However, since the approaches are derived from bio-inspired algorithms the pseudo-code cannot eliminate the terminology surrounding them. Even the original algorithms are slightly verbose for better understanding of the algorithm keeping in mind the biological relevance and analogy.

4. **There are few grammatical mistakes which can be avoided.**
   Response: I have tried my best to rectify the typographic and grammatical errors by doing a thorough proof read of the thesis.

5. **Throughout the thesis, certain information is repeated over and over again. Involved sentences can be rephrased or deleted.**
   Response: I have tried my best to ensure that no information is repeated. Some sentences have been deleted and others have been paraphrased wherever repetition was found.