A
Dissertation
On

Sentiment Analysis using Nature-Inspired Algorithm

Submitted in Partial Fulfillment of the Requirement
For the Award of the Degree of

Master of Technology

in

Software Engineering

by

Shweta Chaudhary
Roll No. 2K14/SWE/17

Under the Esteemed Guidance of

Dr. Akshi Kumar
Assistant Professor

Computer Science & Engineering Department, DTU

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
DELHI TECHNOLOGICAL UNIVERSITY
DELHI - 110042, INDIA
2014-2016

Computer Science & Engineering Department
Delhi Technological University
Delhi-110042
www.dtu.ac.in

CERTIFICATE

This is to certify that the work contained in this dissertation entitle "**Sentiment Analysis using Nature Inspired Algorithms**" submitted in the partial fulfillment, for the award of degree of M.Tech in Software Engineering, Department of Computer Science & Engineering at **Delhi Technological University** by **Shweta Chaudhary**, Roll No. **2K14/SWE/17**, is carried out by her under my supervision. The matter embodied in this project work has not been submitted earlier for the award of any degree or diploma in any university/institution to the best of my knowledge and belief.

Date: ………………

Dr. Akshi Kumar

(Project Guide)

Assistant Professor
Department of Computer Science Engineering
Delhi Technological University

ACKNOWLEDGEMENT

I thank the almighty god and my parents, who are the most graceful and merciful, for their blessing that contributed to the successful completion of this project.

I would like to express deepest gratitude to my guide **Dr. Akshi Kumar** for her full support, expert guidance, understanding and encouragement throughout my study and research. Without her incredible patience and counsel, my thesis work would have been a frustrating and overwhelming pursuit and would never been successful.

I would also like to thank my classmates for their valuable suggestion and helpful discussions.

Shweta Chaudhary

(2K14/SWE/17)

ABSTRACT

The tremendous growth of Web 2.0 has changed the way people express their views and opinions. With the increasing amount of data and information on Web, feature selection is highly essential. As Selecting and extracting feature is itself a exhaustive task that it need to have some automated algorithms to reduce time and space complexity. Traditional techniques for feature selection help reducing feature subset but are of NP hard polynomial nature due to which we need to have some optimized solution. From the past few decades, swarm intelligence is used as optimization techniques for reducing feature subset by decreasing dimensionality and computational complexity resulting in increased accuracy.

In this thesis, we have used Bat Algorithm with SVM for improvement in feature subset with increased accuracy. The algorithm is verified on two different sizes of datasets. Bat algorithm significantly outperformed other algorithms in selecting lower number of features by removing irrelevant, redundant and noisy feature maintaining the accuracy.

Keywords: Sentiment Analysis, Feature Selection, Swarm Intelligence, Bat Algorithm

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1

INTRODUCTION

This chapter briefly introduces the research work proposed in the thesis. Section 1.1 gives an overview of the research undertaken. Section 1.2 discusses motivation and scope. Section 1.3 enlightens the research objectives. Section 1.4 presents an outline of this thesis and labeling the remaining chapters. Finally, Section 1.5 gives the summary of the chapter.

1.1 Introduction

Internet has become an amalgamated, impeccable and a necessary part of our lives. It is changing swiftly so are we. As more and more people have started using it, Web [] is also going through a paramount expedient. In the past few years, web based documents are achieving popularity as a way that portraits individual experience and sentiments. Increment of Web 2.0[] gave rise to applications such as micro-blogging, forums, social-networking sites, wikis etc.

With the growth of Web 2.0[], which emphasis user-generated content, the way people used to express their views and opinions has also changed prominently. Ideas, comments, views, suggestion, feedbacks are shared by the users. Better methods are now used to make decisions. Earlier, people use to conduct surveys but now online reviews are studied to make a conclusion from the opinions given by the user. As with the increase of amount of data on the Web [], it is impossible for an individual to study, examine and extracts sentiments from the large data sets. Hence, one moves to an automated approach which can perform above mentioned process in a promising way, sentiment analysis.

Sentiment Analysis, sometimes referred as opinion mining, used to determine how users are responding for the particular issue. It is the study that determines the attitude of the writer on some documents. It is a kind of text classification which determines the opinions of the writer. Polarity of opinions is classified into three types as positive, negative and neutral. To calculate the percentage of emotions in any comment or view we first need to differentiate essential features from it and then classify it. For selecting a feature, subset is maintained but due to hundreds of thousands datasets the size of search subset is difficult to maintain which leads to redundant data and ambiguity of features leading to a reduced amount of accuracy and precision.

To calculate feature subset some commonly used methods are chi-square, information gain, mutual information etc. are used but they are not successful in reducing corpus size as they produce NP hard nature type problem which is unsolvable in nature. Some more disadvantages related to previously stated methods are discussed later in the chapter. Later to solve the arising problems optimized algorithms are been used in Sentiment Analysis. Algorithms such as Nature-Inspired Algorithms [9], Genetic Algorithms [7], Simulated Annealing [15], etc. are being explored for improving classifier performance. Nature has many sources from which many researchers get inspired and use those algorithms to solve various problems which are difficult to solve or are time complex. Nature-Inspired Algorithms can be classified as Swarm Intelligence Algorithms, Bio-Inspired Algorithms and Physics-Chemistry Algorithms [9]. Due to the dynamic nature of the problem Sentiment analysis has been the extensive area to research. Using Swarm intelligence algorithms helps in local and global patterns for estimating globally best solution.

1.2 Motivation and Scope

Increment of Web 2.0 gives the abundance services which can be helpful for user's awareness. Web 2.0 has involved quite a large number of people to use these services. As almost every type of public is concerned, we need to have refined data which may not offend someone sentiments. So to detect sarcasm and polarity of the documents or of a comment posted online we use automated method. But to the redundancy of data available and noise involved in it we need first optimize feature dataset. Optimization algorithms reduce the noise, redundancy and dimensionality of feature subset resulting in reduced amount of features.

Nature Inspired algorithms has been used as an optimization problems in many area. So the work done in this thesis lay emphasis on Bat Algorithm of selecting only required features. Results and experimental results are shown in forthcoming chapters.

1.3 Research Objectives

The main research objectives of the work done in this thesis are:

Research objective 1 – To study the different area of optimization used in Sentiment Analysis to increase accuracy of results.

Research objective 2 - To propose an optimized Nature inspired algorithm which reduces corpus size and improve complexity problem by reducing noise and ambiguity.

Research objective 3 – To solve high dimensionality of feature subset.

The objective of this thesis is to find an algorithm which can be a hybrid approach to detect polarity of the sentiments with an improve accuracy reducing corpus size.

1.4 Organization of Report

This thesis is structured into 5 Chapters followed by references and appendix.

Chapter 1 presents the research problem, research objectives scope and motivation of the project. Finally, analyzing the need for solution for which research is done.

Chapter 2 provides the essential background and context for this thesis and provides a complete justification for the research undertaken in this thesis.

Chapter 3 gives the details of the methodology employed and outlines the use of Nature-Inspired algorithm in Sentiment Analysis which is proposed approach.

Chapter 4 describes the implementation of algorithm. It discusses all the input sets, platform and tool used to implement result and to compare them.

Chapter 5 describes the experimental results obtained from the given datasets. It presents the analysis of tests performed.

Chapter 6 presents future scope and conclusions based on the contribution made by this thesis.

1.5 Chapter Summary

This chapter presents the idea used in this thesis. It discusses research problem, objectives, goals and motivation for the research. Justification for the research problem is outlined, together with an explanation of the research methodology used. The next chapter describes the literature survey and relevant background work done till date in context of this thesis.

CHAPTER 2

LITERATURE REVIEW

2.1 Internet and Web

Internet is known to be interconnected network which connects different types of systems (laptops, mobile phones, pc etc.) with each other which uses a protocol structure (TCP/IP) for linking these devices. Using Internet one can access different pages which are linked with hypertext documents, e-mails and file sharing.

Internet has become an amalgamated, impeccable and a necessary part of our lives. It is changing swiftly so are we. As more and more people have started using it, web is also going through a paramount expedient. In the past few years, web based documents are achieving popularity as a way that portraits individual experience and sentiments. Web (or "www") is sometimes referred as synonyms of Internet but instead web is just a package which uses Internet to organize itself.

2.2. Evolution of Web

From the last few decades web is rising with interlinked hypertext webpages, images, videos et al. The ongoing web technologies such as HTML, CSS, and XML and so on ensure that all the web based content is supported by all the browsers. The interface between web technologies and browsers helps to build interactive web apps. The concept of web was given by Tim Berners Lee, also known as father of Web, in 1989 in his research. He explained the importance of evolution of web technologies. Some revolutionary ideas were given earlier such as decentralization (no permission required), non-discrimination (Net Neutrality), bottom-up design (integrated code), consensus (transparency) and universality (same language).

Now as with the constant changes in different stages of World Wide Web ("www") Internet has generated a generic platform for different users to provide content. Previously on web 1.0 user could either read or write but not both can be done simultaneously. There was a one way interaction. Web 1.0 was known as “Read OR Write”. Later, after the evolution to Web 2.0 there was a two way communication, i.e. users were able to read and write concurrently. Web 2.0 was known as “Read AND Write”. We are currently at Web 2.0 but there is one more stage which is Web 3.0 and is known as “Semantic Web”. Web 3.0 is yet to be explored. Semantic Web is likely to be as intelligent Web as it might be able to read human minds.

- **Web 1.0 – Read OR Write Web**

It is the traditional World Wide Web. It is called as Read OR Write Web because bloggers write down the content they want to share and publish on static pages. Readers than can read the content and can intuitively conceive their meaning. It could be also said that link is created between readers and writers only when the writer publish their work. As Web 1.0 provides one – way communication it does not facilitate direct or interactive communication.

- **Web 2.0 – Read AND Write Web**

Web 2.0 is referred as social web. It involves number of new generation social technologies like blogs, Wikipedia, social networking sites et al. Here, users are actively involved in team work and communicating two - way communication where not only writers but readers can also read and write to the same web pages. Team work can be defined as a process where different people meet and contact each other to achieve a specific goal or an objective. Sometimes, they interact with other and form an organization. Different people contribute their knowledge and using their skills and experience for development and improvement of some software or a product. Technologies provided by Web 2.0 are:

(i) Blogs

Blog is a descriptive online journal which let user’s post different updates. Post can be in the form of text, pictures, audio or a video file. It is the centerpiece of some mutually discussed topics around the community.

(ii) Wikipedia

Wikis are the collaborative pages of different authors which allow them to post relevant information and links of some famous researches. User can find every detail on that specific topic. Authors on wikis are given separate password to change or to alter the information provided by them.

(iii) Social Networking sites

Social Networking sites like Twitter, Facebook etc. are now a day's use to post regular updates and status of users by setting up a personal user id. They allow users to post photos, videos, feedbacks, comments etc.

- **Web 3.0 – Semantic Web**

It is the third stage of Web evolution and is commonly known as Semantic Web or Intelligent Web as it should be able to read human minds. Currently we exactly do not know what does this of Web evolution will mainly do. Till the point it is only taken as Intelligent Web which will give intelligent responses. It is referred as Read, Write and Request Web. Web 3.0 will not be a simple web page as Web 1.0 nor is it like Web 2.0 which provide style blogs or wiki et al which gives facilities to communicate with human users. Every Web 3.0 will give a little thinking space; each user on Web 3.0 will be a reader, writer and a requester/execute simultaneously.

Semantic here means data driven which implies of bridging the communication space between human mind and online applications. Taking an example for the above line when a user searches more about "online electronic products" he will be receiving latest updates on it via advertisements. Later, if the same user refines his/her search to "television and cellular phones"; Web will know that user is planning to buy "television and cellular phone online"; so Web will give an automated search that combine both the queries.

Web Evolution can be summarized as connecting real minds to World Wide Web in Web 1.0; Web 2.0 ensures real minds using World Wide Web and Web 3.0 will provide a virtual environment of real minds that use World Wide Web.

2.3. Sentiment Analysis

Sentiment Analysis, sometimes referred as opinion mining, used to determine how users are responding for the particular issue. It is the study that determines the attitude of the writer on some documents [1]. It is a kind of text classification which determines the opinions of the writer. With Web 2.0, which emphasizes user-generated content, the way people use it to express their views and opinions have also changed prominently. Ideas, comments, views, suggestion, feedbacks are shared by the users which assist in decision-making. Earlier people used to conduct surveys but now online reviews are studied to make a conclusion from the opinions given by the user. As with the increase of amount of data on the Web, it is impossible for an individual to study, examine and extract sentiments from the large data sets. Hence, one moves to an automated approach which can perform above mentioned process in a promising way, sentiment analysis.

Sentiment Analysis or Opinion Mining deals with the sentiment or opinions given by users online in either a review or on social networking sites [1]. Some users while purchasing goods needs to decide among several products with same features, so they depend on other users to buy products they are using. Sentiment lexicon also called opinion word plays very important role in Sentiment Analysis. Generally, adjective followed or preceded by adverb is detected as opinion word or phrase. To detect the adverb effect on the sentence we use tagging. Tagging can be done using POS taggers (**Part-of-speech taggers).** It detects adjective and nearest adverb in a sentence. Sometimes it is quite easy to detect sarcasm as some reviews show directly negative or positive nature such as good, bad etc. [2]. But sometimes these words are so difficult to distinguish as with a negation before them they changes their meaning like “not so good in taste”, here good can be considered as a positive word but with a negation preceding, the meaning can completely changes. The most stimulating task is to detect polarity of those sentences that don’t have any adjective preceded by them. This is a puzzling task in any area of research. Sentence Level Sentiment Analysis is comparatively quite easy. If there is a direct sentiment attached it become tranquil to detect polarities and sentiments. But this is also not true in every case as if there is no direct sentiment or phrase that gives polarity, we need to analyze whole sentence to detect sentiment from it. It becomes difficult for a machine to analyze it.

Document Level sentiment analysis is the toughest things to do as whole opinion in a document is considered to a single module.

2.3.1. Level of Granularity

Extraction of opinions can be done at several levels of granularity. Above figure shows the levels of sentiment extraction. The sentiment analysis tasks can be done in following levels of granularity:

- Word Level Sentiment Analysis
- Sentence Level Sentiment Analysis
- Document Level Sentiment Analysis
- Feature Based Sentiment Analysis

2.3.2. Challenges

- **Sarcasm**

Detecting sarcasm is also difficult to do as there can be reviews which use positive words but their meaning is different, like, “What a great online shopping site, can’t find anything useful”. “Great” is used but whole meaning of the sentence is different and is a negative one. This type of problem is a part of sentiment analysis subjectivity.

- **Comparison**

Comparative opinion detection is a field of study in Sentiment Analysis, “Brand A phone is better than Brand B phone”, this review has no sentiment in positive or negative sense. It just shows Brand A is better Brand B.

- **Spam Detection**

Detecting spam is also difficult because sometimes representatives of that product duplicates so many fake reviews or their competitors provides negative reviews then it becomes hard to design algorithms for such problems.

2.3.3. Sentiment Analysis Tasks

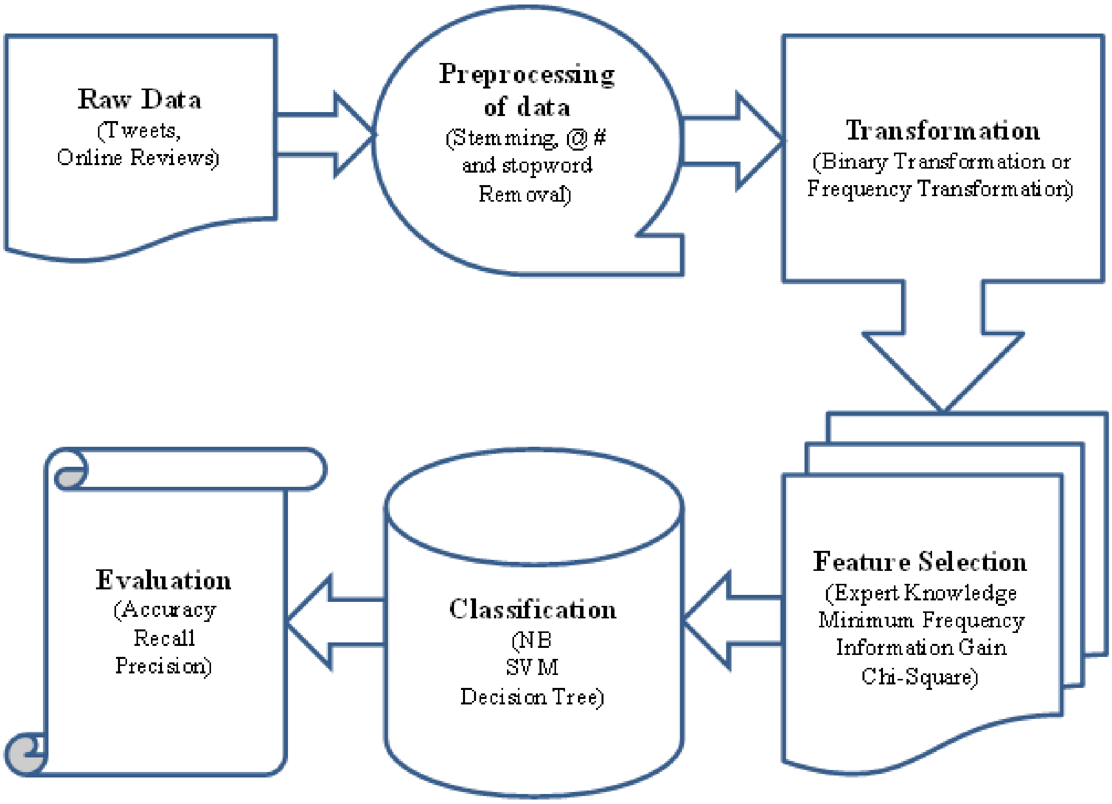


Figure 2.1 Sentiment analysis framework

a) Data Source

User's opinion is a vital part for advancement of any product for better results. Opinions can be collected from various sources such as social media sites, Blogs, Audit sites etc.

- **Social Media sites**

Social Media sites also termed as micro blogging sites allow user to connect through the entire world virtually by which users can share or express their views. Twitter, Facebook are most popular social media sites.

- **Blogs**

Blogs can be defined as website which is updated on frequent intervals (usually on daily basis). These blogs are typically run by small group of particular organization or individuals. They share their personal experience, feeling and emotions about any products or aftereffect of some issues.

- **Audit Sites**

While purchasing a product or choosing from different products other opinions plays an important role. People use the product and give their views on the same which helps other for making their decision. One of the most famous website which allows users to have reviews is www.consumerreports.com . Apart from this there are many e-commerce websites sites such as www.flipkart.com (product review), www.tripadvisor.com (travel review).

b) Preprocessing of Data

After data collection it has to be preprocessed. Preprocessing of data means converting incomplete data into accessible format for classification. Data collected from different sources are either incomplete or have many errors which has to be removed before feature selection. The following steps are used to extract the main content from the data set-

- **Data Cleaning**

i) Removal of urls
ii) Removal of hashtags (#)
iii) Removal of quotes (@tags)
iv) Removal of punctuation marks (,.?)
v) Remove repeated words
vi) Remove Smiley Emoticons
vii) Remove special Symbols
viii) Remove ‘Wh’ questions

- **Data Transformation**

i) Tokenization

Tokenization is the process of dividing the text into words or phrases which can be termed as list of tokens. A word tokenizer algorithm is used to estimate the occurrence of particular word or a phrase in a sentence. It keeps a count of repeating words and then reducing it to one word. It can be represented in vector space as

Data = (word, frequency, polarity)

ii) Stop words Removal

There are many words which occur frequently in a document but there uses are meaningless as they do not contribute in the polarity of the statement. They are mostly used to combine words. Example of stop words are ‘is’, ‘are’, ‘and’, ‘hence’, ‘an’ etc. These words must be removed before classification.

iii) Stemming

Stemming is the process of reducing the words to its base or root form. Example the word: ‘ate’, ‘eaten’, ‘eating’ can be replaced to ‘Eat’.

iv) Handling Negation

Negation handling is a difficult job as, “I don’t want to eat” and “I want to eat” is different only by one word ‘not’ but both the sentence is of opposite polarity.

c) Feature Selection

Next step in Sentiment Analysis is feature selection. It reduces the amount of data for analysis and making it more potent. Previously used techniques for feature selection are document frequency (Bai, 2011; Dang et al., 2010; Pang et al., 2002), mutual information (Li et al., 2009; Turney, 2002), information gain (Abbasi et al., 2011, 2008; Li et al., 2009; Riloff, Patwardhan, & Wiebe, 2006) and chi-square (Abbasi et al., 2011; Li et al., 2009). Although except Information gain no other algorithm is accepted as a universal method for feature selection [3]. But due to the corpus size Information gain is also not used extensively. To solve the corpus problem and to reduce the high dimensionality problem we in this study discuss an optimization algorithm [6]. Later in this chapter all the other problems related to selecting feature has been discussed with the novel method.

d) Classification

After feature selection step we will have enhance features, which will be the input to Text Classification step. For classification we have different machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), Decision Tree and many more. These machine learning algorithms will read the features extracted from previous step and will give the output as positive, neutral or negative. Although selecting the best machine learning algorithm for classification depends on the types of requirement and accuracy.

Classification Techniques

Machine classification learning usually distinguishes between three learning methods: supervised, weakly supervised and unsupervised learning.

- **Supervised Learning**

Supervised machine learning techniques associate the use of a labelled training dataset to learn a certain classification function and involve learning a function from examples of its inputs and outputs. The output of this function is either a continuous value (‘regression’) or can predict a category or label of the input object (‘classification’). In this section different machine classification algorithms will be discussed.

- **Naïve Bayes Classifier**

Naïve Bayes classifier takes assumption to calculate probabilities to classify documents. To calculate entire probability, multiplies probabilities of individual’s works in those documents because it requires less computational time. NB classifier is used as a baseline method as it gives results which are sufficiently good as compared to other method [43].

- **Maximum Entropy Classifier**

Maximum entropy combines joint features that are generated from the set of features by encoding. Encoding mapping forms a feature set and a label to a vector. It is also known as the exponential or log-linear classifier because they work by extracting some set of features from input domain combining them linearly and then using this sum as an exponential function [44].

- **Decision Tree (DT)**

Decision Tree classifier is a tree in which all internal nodes are considered as features, edges which have single node left are labelled as feature weight. It classifies the document by starting from the tree node and moving successfully in downward from the branches until a leaf node is encountered. Document is then classified into the category of leaf nodes [45].

- **Support Vector Machines (SVMs)**

SVM are considering to yields highest accuracy results in classification. Support vector machines (SVM), a discriminative classifier is considered the best text classification method [46]. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which Multi class SVM is used for Sentiment classification [47].

Other supervised learning methods include Decision Rule classifiers, (Artificial) Neural Networks, Logistic Regression, Rocchio Methods and Random Forests.

- **Weakly-Supervised and Unsupervised Learning**

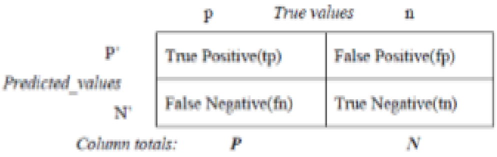
Due to unavailability of labelled corpus dataset is not always possible to use supervised methods. Weakly and unsupervised methods then are opted for machine learning that does not require part-of-speech tagging data.

All the learning patterns in the input where no specific output value are supplied are distinguished as unsupervised methods. In this learner only receives some unlabeled data. Examples of unsupervised learning methods are (k-means) clustering or cluster analysis, the problem of discerning multiple categories in a collection of objects and the expectation-maximization algorithm, an algorithm for finding the maximum likelihood of examples [48].

Weakly-supervised learning, or semi-supervised learning, involves learning a classification task from a small set of labelled data and a large set of unlabeled data [49].

e) Interpretation

Various methods are used to measure the performance of classifiers like calculating precision, accuracy, recall and F-measure. All the measures can be calculated using confusion matrix. Confusion matrix is shown in fig.as.



		p	True values	n
Predicted_values	P'	True Positive(tp)		False Positive(fp)
	N'	False Negative(fn)		True Negative(tn)
Column totals:		P		N

Depending on the polarity (negative or positive) results are divided into two main classes positive (P) and negative (N). We assume that classifier output can have mainly four outcomes: true positive (tp), true negative (tn), false positive (fp), false negative (fn). Matrices can be calculated as:

1. Accuracy = (tp+tn)/(P+N)

2. Precision = tp/(tp+fp)

3. Recall = tp/P

4. F-measure = 2/ ((1/precision)+(1/recall))

2.4. Feature Selection

Literature studies have revealed that the sentiments have been typically researched at the document level, sentence level, entity and feature or aspect level. Given an opinion document *d*, we can discover all opinion quintuples (e_i, a_{ij}, s_{ijkl}, h_k, t_l) in *d*, where e_i is a unique entity, a_{ij} is the unique aspect for entity e_i, s_{ijkl} defines the sentiment, i.e., it classifies the sentiment into positive, negative or neutral categories.

The fourth component and fifth components are opinion holder and time respectively. They can also be used for extracting and categorizing entities and aspects. Aspect or feature selection is identified as one of the most significant responsibility in sentiment analysis. In feature selection, the feature is first identified, followed by the selection procedure and then extraction and reduction process if required. Feature identification includes comprehending feature types such as term frequency, term Co-occurrence, Part-of-Speech and Opinion Words, for identification purpose. Numerous methodologies are applied to the problem of feature selection in text categorization. The major and frequently used approaches are document frequency (DF), information gain (IG), mutual information (MI), x^2-test (CHI) and term strength (TS). The CHI square and information gain are more impressive for the conclusion of optimized classification results, and if a small degeneration in effectiveness is nominal than document frequency is an enhanced choice for efficiency and scalability [3].

2.4.1. Advantages and uses

Feature Selection in sentiment analysis is undertaking a variety of subjects such as huge feature space problems, repetition, domain dependency, difficulty in implicit feature identification, and limited work on Lexicon-structural features, amongst others [4]. The dynamic objections of feature extraction have been acknowledged across published research as high dimensionality, redundancy [6], Domain dependency [8], Lexicon-structural features, POS tagging, amongst others. The main goal of Feature Selection is to minimize feature subset with improved classification accuracy. It only selects valuable features and eliminating irrelevant, redundant, or noisy features and thus sinking the number of features. While doing this size of feature set increases in terms of complexity. The main challenge in feature selection is to maintain the corpus size without compromising the accuracy. Hence, we need to select some representative features from the original feature space to reduce the dimensionality of feature space and improve the efficiency and performance of classifier.

Generally, the numbers of words are marked as candidate features, most of them are stop words which have no meaning and they don’t contribute in polarity of the sentence. These meaningless words are removed as they only decrease classifier accuracy. So, feature selection involves optimal searching using some specified methods with exhaustive and approximate method. Exhaustive search method is helpful and also gives optimal solution to some extend but it don't work on large datasets. To find an optimal subset solution is an NP-hard problem, for large number of features. If there is N number of features then the possible output would be exponential to 2^n. Due to the non-practical implementation of exhaustive search, the centres have moved to meta-heuristic approaches (approximate methods subclasses) from search strategies. Due to dynamic nature of problem and to achieve fruitful classification accuracy some successful research are done in past and are also currently being explored.

2.4.2. Goals of feature selection

Feature Selection is considered as a most important step in tasks of Sentiment Analysis. The amount of data collected causes problem in construction of classification model as it takes large memory or time taken by it would make it a NP Hard problem. Creating a feature subset helps us to work within the designed and assumed constraints. If the number of features is reduced in a manner that does not significantly increase accuracy we would get a smaller model. This would make the model more unreadable and it could also be less over-fitted. Removing redundant and irrelevant features can help improve the performance of classifiers. This is done by reducing the potential for over-fitting and, in the case of redundant features, selecting those that work best.

2.4.3. Review on Previous Algorithm for Feature Selection

Currently feature selection methods uses numerical value to assign features based on statistical equation and using these values appropriate features are selected from the sorted feature vector. Selection of threshold value is user dependent and impacts the classification accuracy. This results in selection of sub-optimal feature set and thereby consuming more processing power and more resources. Some traditional methods are Information Gain, Documents Frequency thresholding and Chi Square.

Selecting feature subset from high dimensional feature space is labeled as globally optimized problem which aims at reducing irrelevant and noisy features and decreasing complexity problem.

2.5. Nature-inspired Optimization Algorithm

Nature has a rich source of ideas from which many researchers are being inspired. Today, in almost every field Nature-Inspired Algorithms are used to have an optimized solution for a problem. Nature-Inspired Algorithms can be classified as Swarm Intelligence Algorithms, Bio-Inspired Algorithms and Physics-Chemistry Algorithms [**9**].

Swarm intelligence is collection of some local agents of same community who follows same and simple procedures to communicate with each other in their local environment. They don't have any central supervisor instead it is just collective behavior of animals, small insects or some naturally occurring phenomenon that help each other in either static or a dynamic manner.

Swarm Intelligence based algorithms

Swarm intelligence involves simple agents who interacts each other locally in their respective environment. Some nature designed rules are followed to exchange information between agents within the environment. These agents works unintelligent way but concluding the system gives a new dimension to the whole system. They follow simple principle of decentralization and self-organization of group of interacting agents from which a global intelligent behavior is recognized. A number of algorithms have been researched and are extensively used in optimization. Some Swarm intelligence algorithms are shown in Table 1. They show remarkable results in NP-Hard problems and generate better solutions. Due to the scalability and strength of SI-based algorithms they have become the foremost choice in judgment of solutions for optimization problems.

Reference	Algorithm	Author	Year
[25]	Ant Colony Optimization	M Dorigo	1992
[28]	Artificial bee colony	Dervis Karaboga	2005
[16]	Bacterial foraging	Kevin M Passino	2002

[38]	Bat algorithm	Xin-She Yang	2010
[29]	Bee colony optimization	Dusˇan Teodorovic´, Mauro Dell’Orco	2005
[24]	Bee Hive	H.F. Wedde, M. Farooq,Y. Zhang	2004
[26]	Bee system	P Lucic , D Teodorovic	2001
[22]	Bees algorithms	DT Pham, A Ghanbarzadeh, E Koc, S Otri, S Rahim, M Zaidi	2006
[30]	Bees swarm optimization	Habiba Drias, Souhila Sadeg, Safa Yahi	2005
[42]	BumbleBees	Francesc Padro´, Jesu´s Navarro	2011
[39]	Consulted Guided Search	Serban Iordache	2010
[11][12]	Cuckoo search	Xin-She Yang, Suash Deb	2009
[40]	Eagle Strategy	X. S. Yang and S. Deb	2010
[36]	Fast bacterial swarming algorithm	Ying Chu, Hua Mi, Huilian Liao, Zhen Ji, QH Wu	2008
[37]	Firefly Algorithm	X.-S. Yang	2008
[18]	Fish Algorithm	X.-L. Li, Z.-J. Shao, J.-X. Qian	2002
[34]	Glowworm swarm optimization	KN Krishnanand , D Ghose	2005
[33]	Good lattice swarm optimization	Shoubao Su, Jiwen Wang, Wangkang Fan, and Xibing Yin	2007
[14]	Krill Herd Algorithm	Amir Hossein Gandomi, Amir Alavi	2012
[35]	Monkey Search	Antonio Mucherino, Onur Seref	2007
[20]	Particle swarm optimization	Dr. Eberhart , Dr. Kennedy	1995

[17]	Virtual ant algorithm	X-S Yang, J M Lees, C T Morley	2006
[31]	Virtual Bees	X.-S. Yang	2005
[41]	Weightless Swarm Algorithm	To Ting, Ka Lok Man, Sheng-Uei Guan, Mohamed Nayel, and Kaiyu Wan	2010
[15]	Wolf search	Rui Tang, S. Fong, Xin-She, S. Deb	2012

TABLE 2.1 Swarm Intelligence Algorithms

2.6 Feature Selection Approaches

There are primarily two approaches used for selecting features. First approach is wrapper and filter approach and second approach is evolutionary approach [50].

2.6.1 Wrapper and filter approach

Wrapper method - As the name tells, this approach “wraps around” the induction algorithm which is then used for final classifier. This approach follows a heuristic search with backward or in forward direction. In general, wrapper method is quite costly to perform as it consists of N-fold cross-validation for accuracy calculation. As it selects the whole set of variables instead of one variable at a time, this characteristic makes it a helpful approach.

Filter method – This method make selection based on statistics only. It only selects those variables which has highest information gain. Basically, it is divided into two categories; one which evaluates all features in one go and other which evaluates features in multiple combined sets with heuristic search.

Another approach which combines wrapper and filter method for searching is considered as hybrid feature selection method.

2.6.2 Evolutionary approaches

As selecting feature subset is NP-hard problem and are mostly unsolvable numbers of different evolutionary approaches [10] are proposed for it. Broadly evaluated area is genetic algorithms. Recently evaluated area for selecting feature selection is Nature – Inspired algorithm using them as optimization techniques. Some algorithm used for solving the problem are ant colony optimization, particle swarm optimization etc.

2.7. Feature Selection using Swarm Intelligence

Swarm intelligence algorithms are used in extensive area where we keen to find the optimized results. These algorithms help in reducing feature subset by working in reduced number of iterations till best result is not attained. Swarm intelligence helps in improving superiority of solution set by working in iterations until best result is not obtained. Reduced features in feature set increments accuracy in just four simple steps including generation of feature subset, evaluation of subset, termination condition and validation of results.

Starting from generating feature subset the local feature subset is searched. Depending on some search criteria local feature sets are evaluated and compared with best search value which has been previously evaluated. If in case current solution is better than the previous best solution it gets replaced to best solution. Generation of feature subset continues till stopping condition is attained. The stopping criterion depends on error rate and number of iterations to be done. When a certain amount of threshold is above from error rate then the algorithm stops or if number of iteration goes beyond the specified number of cycles.

Table 2 shows the comparisons of various Nature-Inspired algorithms which are used for Sentiment Analysis with their accuracy results. Using these algorithms gives promising accuracies with reduced feature set. Experiments have shown that reducing the number of features up to 36%, it is possible to preserve an accuracy of 87.15%. [12]. According to the survey done, it shows that ABC and PSO are the most powerful optimization techniques for solving hybrid problems. These methods have been used for optimizing the feature subset selection successfully by researchers and have improved the accuracy of classification as stated in table 2. Although we have a variety of nature inspired algorithms for optimization as can be seen from table 1 only few have been explored in domain of opinion mining. The results obtained using Nature – Inspired Algorithms prove to be better in terms of accuracy. As swarm intelligence in sentiment analysis is a current research area many researchers are investing in this area.

SI technique	Data set	Classifier	Accuracy without Optimization	Accuracy with Optimization	Year	Reference
PSO	Twitter Data	SVM	71.87	77	2012	[27]
hybrid PSO/ACO 2	Product Reviews, Governmental decisions data	Decision Tree	83.66	90.59	2014	[21]
Artificial Bee Colony	Internet Movie Database (IMDb)	Naïve Bayes	85.25	88.5	2014	[23]
		FURIA	76	78.5		
		RIDOR	92.25	93.75		
Ant-Bee Colony	Product Reviews	SVM	55	70	2015	[19]
PSO	Restaurant Review Data	CRF	77.42	78.48	2015	[32]

TABLE 2.2 Comparisons of Various Swarm Intelligence Techniques on Sentiment Analysis

2.8. Bat Algorithm

2.8.1. Behavior of Bats

Bats are captivating mammals that uses echolocation for finding its prey. As and when certain stop criteria are met selection of best search stops. The dynamic behaviors of bats are used to control the balance between exploitation and exploration. Bats use a type of sonar, called, echolocation, to detect prey, avoid obstacles, and locate their roosting crevices in the dark. These bats emit a very loud sound pulse and listen for the echo that bounces back from the surrounding objects [55][56]. Their pulses vary differently which is dependent on different type of species.

2.8.2. Sound quality of Echolocation

Though each pulse only lasts a few thousandths of a second (up to about 8 to 10 ms), however, it has a constant frequency which is usually in the region of 25kHz to 150 kHz. The typical range of frequencies for most bat species are in the region between 25kHz and 100kHz, though some species can emit higher frequencies up to 150 kHz. Each ultrasonic burst may last typically 5 to 20 ms, and microbats emit about 10 to 20 such sound bursts every second. When hunting for prey, the rate of pulse emission can be sped up to about 200 pulses per second when they fly near their prey. Such short sound bursts imply the fantastic ability of the signal processing power of bats. In fact, studies shows the integration time of the bat ear is typically about 300 to 400 µs. As the speed of sound in air is typically $v = 340$ m/s, the wavelength λ of the ultrasonic sound bursts with a constant frequency f is given by

$$\lambda = v/f$$

This is in the range of 2mm to 14mm for the typical frequency range from 25 kHz to 150 kHz. Such wavelengths are in the same order of their prey sizes.

Amazingly, the emitted pulse could be as loud as 110 dB, and, fortunately, they are in the ultrasonic region. The loudness also varies from the loudest when searching for prey and to a quieter base when homing towards the prey. The travelling range of such short pulses is typically a few meters, depending on the actual frequencies. Bats can manage to avoid obstacles as small as thin human hairs. Studies show that bats use the time delay from the emission and detection of the echo, the time difference between their two ears, and the loudness variations of the echoes to build up three dimensional scenario of the surrounding. They can detect the distance and orientation of the target, the type of prey, and even the moving speed of the prey such as small insects. Indeed, studies suggested that bats seem to be able to discriminate targets by the variations of the Doppler effect induced by the wing-flutter rates of the target insects. Obviously, some bats have good eyesight, and most bats also have very sensitive smell sense. In reality, they will use all the senses as a combination to maximize the efficient detection of prey and smooth navigation. However, here we are only interested in the echolocation and the associated behavior. Such echolocation behavior of bats can be formulated in such a way that it can be associated with the objective function to be optimized, and this makes it possible to formulate new optimization algorithms.

2.8.3. Initial condition for Bat Algorithm

If we idealize some of the echolocation characteristics of bats, we can develop various bat-inspired algorithms or bat algorithms. In the basic bat algorithm developed by Xin-She Yang (2010), the following approximate or idealized rules were used.

1. All bats use echolocation to sense distance, and they also 'know' the difference between food/prey and background barriers in some magical way.

2. Bats fly randomly with velocity v_i at position x_i with a frequency f_{min}, varying wavelength and loudness A_0 to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $r \in [0, 1]$, depending on the proximity of their target.

3. Although the loudness can vary in many ways, we assume that the loudness varies from a large (positive) A_0 to a minimum constant value A_{min} .

```
Initial Conditions:
    Objective function $(x)$, $x = (x^1 \ldots x^n)$.
    Initialize the bat population $x_i$ and $v_i$, $i = 1, 2 \ldots m$.
    Define pulse frequency $f_i$ at $x_i$, $\forall i = 1, 2 \ldots m$.
    Initialize pulse rates $r_i$ and the loudness $A_i$, $i = 1, 2 \ldots m$.
While $t$<T
    For each bat $bi$, do
        Generate new solutions using Equations (1), (2) and (3).
        If $rand > ri$, then
            Select a solution among the best solutions.
            Generate a local solution around the best solution.
        If $rand < Ai$ and $(xi) < (\hat{}x)$, then
            Accept the new solutions.
            Increase $ri$ and reduce $Ai$.
Rank the bats and find the current best $x\hat{}$.
```

Table 2.3: Pseudo Code for Bat algorithm

2.8.4. Binary Bat Algorithm

Mirjalili et al. in 2014 proposed Binary Bat Algorithm in 2014 [51]. A binary search space can be considered as a hypercube. The search agents (particles) of a binary optimization algorithm can only shift to nearer and farther corner of this hypercube by flipping various numbers of bits. Hence, in designing the binary version of BA, some basic concepts of the velocity and position updating process must be modified. In the continuous version of BA, the artificial bats can move around the search space utilizing position and velocity vectors (or updated position vectors) within the continuous real domain. Consequently, the concept of position updating can be easily implemented for bats by adding velocities to positions. However, the meaning of position updating is different in a discrete binary space. In a binary space, due to dealing with only two numbers (‘‘0’’ and ‘‘1’’), the position updating process cannot be done. Therefore, a way must be found to use velocities for changing agents’ positions from ‘‘0’’ to ‘‘1’’ or vice versa. In other words, a link has to be devised between velocity and position as well as revising the position. In discrete binary spaces, position updating means switching between ‘‘0’’ and ‘‘1’’ values. This switching should be done based on the velocities of agents [52]. The question here is how the concept of velocity in a real space should be employed in order to update the positions in a binary space. The idea is to change the position of an agent with the probability of its velocity. In order to do this, a transfer function is necessary to map velocity values to probability values for updating the positions. In other words, a transfer function defines the probability of changing a position vector’s elements from 0 to 1 and vice versa. Needless to say, transfer functions force particles to move in a binary space. To solve BBA for solving binary problems a V-shaped transfer function has been opted.

According to Rashedi et al.the following concepts should be taken into account for selecting a transfer function in order to map velocity values to probability values.

- The range of a transfer function should be bounded in the interval [0, 1], as they represent the probability that a particle should change its position. A transfer function should provide a high probability of changing the position for a large absolute value of the velocity. Particles having large absolute values for their velocities are probably far from the best solution, so they should switch their positions in the next iteration.

- A transfer function should also present a small probability of changing the position for a small absolute value of t velocity.

- The return value of a transfer function should increase as the velocity rises. Particles that are moving away from the best solution should have a higher probability of changing their position vectors in order to return to their previous positions.

- The return value of a transfer function should decrease as the velocity reduces.

These concepts guarantee that a transfer function is able to map the process of search in a continuous search space to a binary search space while preserving similar concepts of the search for a particular evolutionary algorithm. The transfer function that has been used for binary PSO is presented as Eq. (1) [56].

$$S\left(v_i^k(t)\right) = \frac{1}{1 + e^{-v_i^k(t)}} \tag{1}$$

Where $v^k_i(t)$ the velocity of particle i in k-th dimension at iteration t. After calculating the probabilities using transfer functions, a new position updating equation is necessary to update particles' position as follows [56]:

$$x_i^k(t+1) = \begin{cases} 0 & \text{If} \quad \text{rand} < S\left(v_i^k(t+1)\right) \\ 1 & \text{If} \quad \text{rand} \geq S\left(v_i^k(t+1)\right) \end{cases} \tag{2}$$

Where $x^k_i(t)$ and $v^k_i(t)$ indicate the position and velocity of i-th particle at iteration t in k-th dimension. This method has a drawback as the particles are forced to take values of 0 or 1 [56]. So the particles remain unchanged in their positions when their velocity values increase. However, according to the concepts mentioned above for designing a transfer function, a better way is to indulge the particles with high velocity to switch their positions. A v-shaped transfer function and position updating rule are proposed in order to do this as in Eq. (3) and (4)

$$V\left(v_i^k(t)\right) = \left|\frac{2}{\pi} \arctan\left(\frac{\pi}{2} v_i^k(t)\right)\right| \tag{3}$$

$$x_i^k(t+1) = \begin{cases} \left(x_i^k(t)\right)^{-1} & \text{If} \quad \text{rand} < V\left(v_i^k(t+1)\right) \\ x_i^k(t) & \quad \text{rand} \geq V\left(v_i^k(t+1)\right) \end{cases} \tag{4}$$

where $x^k_i(t)$ and $v^k_i(t)$ indicate the position and velocity of i-th particle at iteration t in k-th dimension, and $(x^k_i(t))^{-1}$ is the complement of $x^k_i(t)$. The steps of utilizing the proposed transfer function to force particles to move in a binary search space.

Eq. (3) is employed as the transfer function in order to map the velocities of BBA's bats to the probabilities of flipping their position vectors' elements. Consequently, the rules of (4) are used to update position vectors.

```
Initialize the bat population: Xi (i = 1, 2... n) =rand (0 or 1) and $V_i$=0
Define pulse frequency $F_i$
Initialize pulse rates $r_i$ and the loudness $A_i$
while (t < Max number of iterations)
    Adjusting frequency and updating velocities
    Calculate transfer function value using equation (3)
    Update positions using equation (4)
    if (rand > ri)
        Select a solution (Gbest) among the best solutions randomly
        Change some of the dimensions of position vector with some of the
        dimensions of Gbest
    end if
    Generate a new solution by flying randomly
    if (rand < Ai & f(xi) < f(Gbest))
        Accept the new solutions
        Increase ri and reduce Ai
    end if
    Rank the bats and find the current Gbest
end while
```

Table 2.4: Pseudo Code for Binary Bat algorithm

2.8.5. Advantages of using Bat Algorithm

Bat algorithm has been used in many problems as an optimized algorithm and it has shown outstanding results [53]. A detailed comparison of bat algorithm (BA) with genetic algorithm (GA), PSO and other methods for training feed forward neural networks concluded clearly that Bat Algorithm has advantages over other algorithms [54].

2.9. Related Work

In terms of efficiency and complexity, Swarm inspired algorithms are been practiced more as compared to other feature selection method to have accurate results without compromising with other dimensions.

CHAPTER 3

PROPOSED WORK

Chapter 2 identified number of issues related to feature selection in Sentiment Analysis using traditional method. This chapter illustrates a novel approach that helps in achieving better results (in terms of accuracy, precision and recall) as compared to previously used algorithms. Section 3.1 gives an overview of the research undertaken. Section 3.2 portraits the architectural view of the proposed paradigm. Section 3.3 describes each module of the system and how proposed algorithm works. Lastly, Section 3.4 gives the summary of the chapter.

3.1 Proposed Framework

Due to the increase in Web 2.0 services, the amount of data available online has changed drastically in terms of volume as it has become a global source of useful information. Sentiment Analysis works on finding opinions of different minds; Natural Processing cannot deals with the corpus size of the dataset. Section introduces a number of automated approaches such as Information gain [], Chi Square [] et al for selecting feature subset. These methods assign a numerical value on features selected from a semi sorted feature vector. Selecting features gives a range of threshold values from which user selects the appropriate value according to the classification accuracy. Furthermore, obtained feature set is then fragmented into sub-optimal feature set which results in unbearable use of resources and processing time.

Selecting feature subset is a NP hard problem as reducing high dimensional feature space is an optimizing problem. An optimization algorithm removes redundancy, extraneous and noisy features yielding accurate classification and reduction in processing time. Swarm Intelligence algorithms are broadly used as an optimization algorithm. Table in Section 2 shows some Swarm Intelligence Algorithms which are used till date for optimizing feature subset in Sentiment Analysis. In this research, we use Bat Algorithm as an optimization algorithm and the outcomes are then compared with results obtained using WEKA Tool on the same Twitter Dataset with SVM (Support Vector Machine) classifier.

3.2 Architectural View

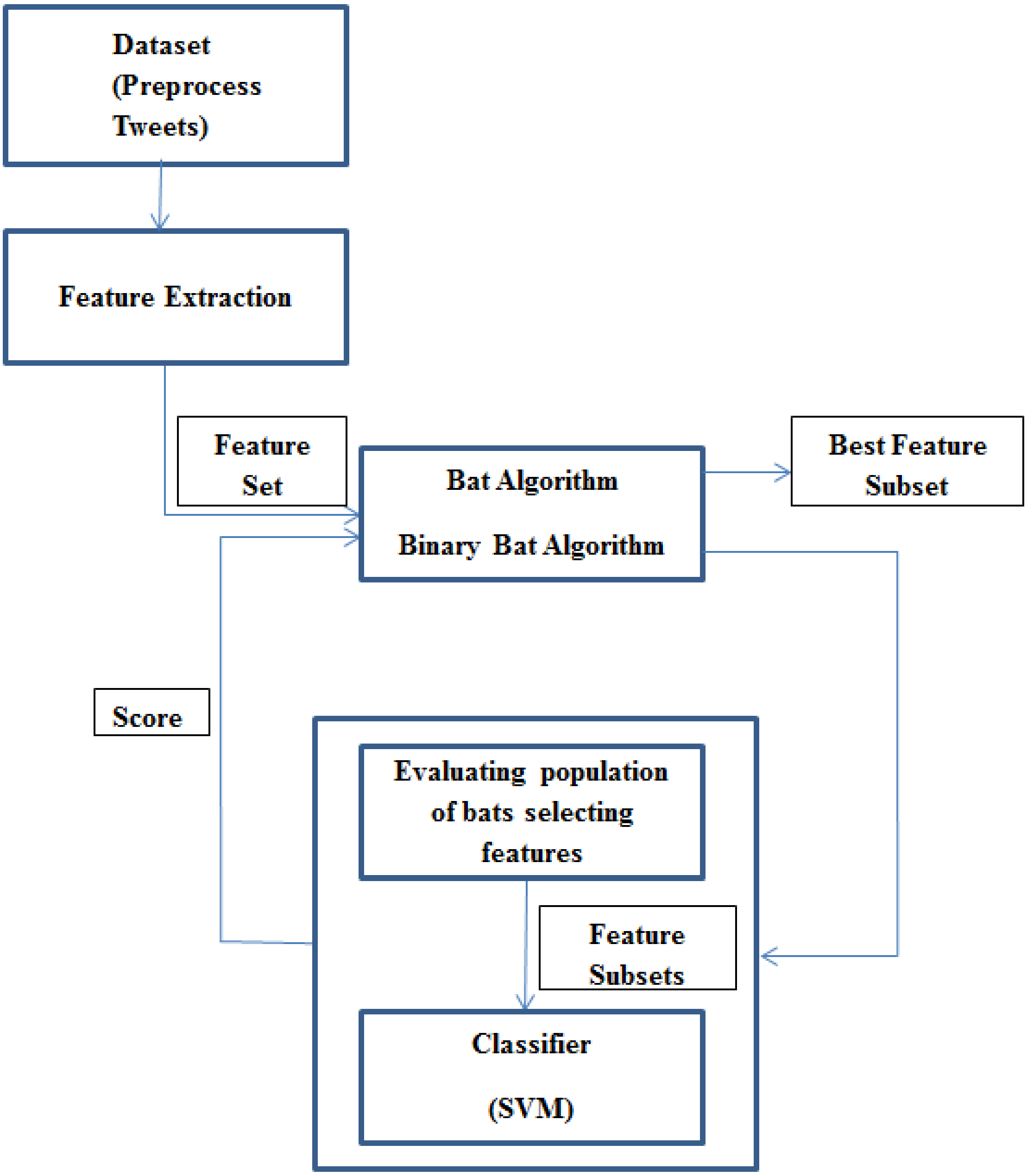
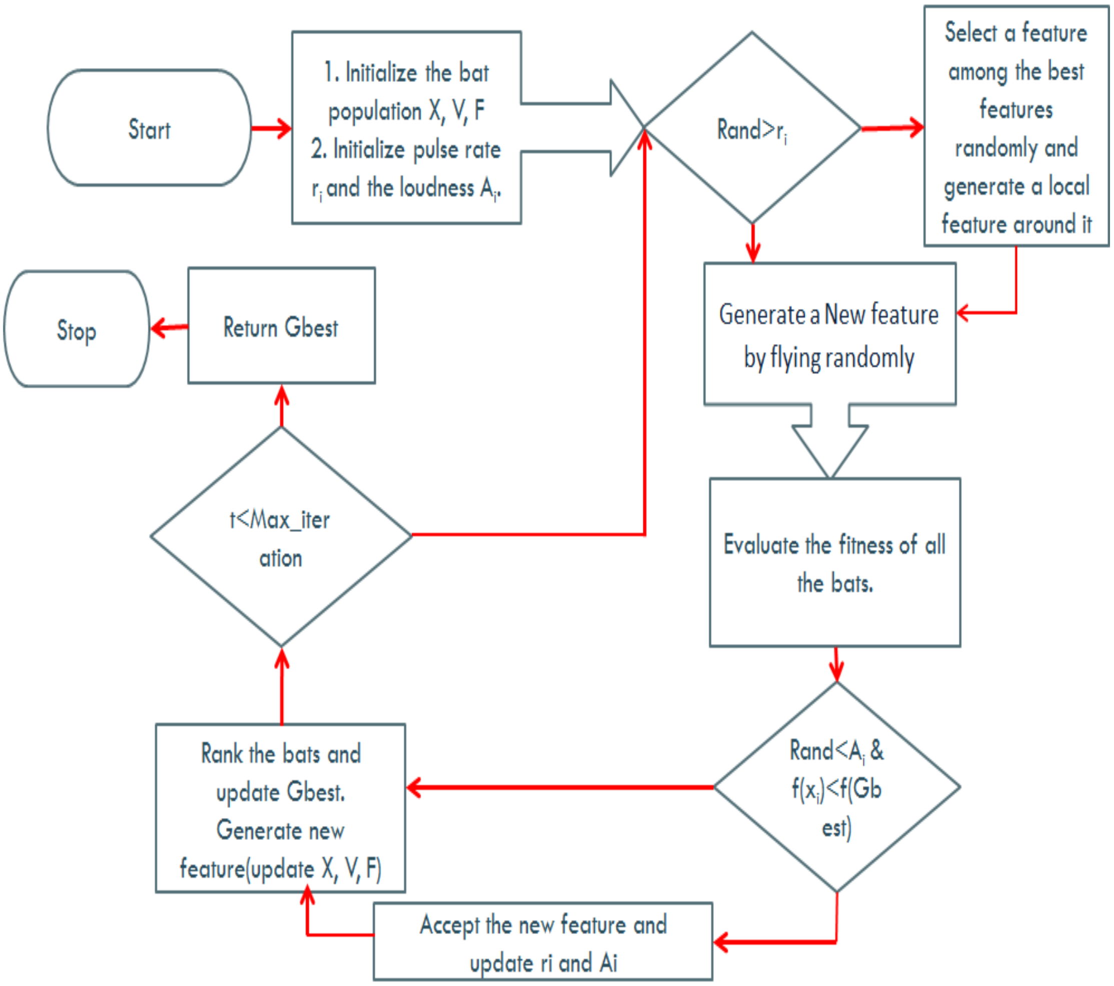


Figure 3.1: Pictorial view for using Bat algorithm

The proposed approach firstly retrieves tweets from Twitter and then preprocess them giving feature subsets for classification. For feature classification instead of using

Traditional Methods; Bat Algorithm and Binary Bat Algorithm are used to improve high dimensionality problem. Later, SVM classifies the tweets into positive, negative and neutral behavior of opinion holder. Also, training data is maintained for calculating accuracy, precision and recall. Figure shows the overview of the system proposed in this research.

Working of Bat Algorithm



Start
1. Initialize the bat population X, V, F
2. Initialize pulse rate r_i and the loudness A_i.
Rand>r_i
Select a feature among the best features randomly and generate a local feature around it
Stop
Return Gbest
Generate a New feature by flying randomly
t<Max_iteration
Evaluate the fitness of all the bats.
Rank the bats and update Gbest. Generate new feature(update X, V, F)
Rand<A_i & $f(x_i)$<f(Gbest)
Accept the new feature and update ri and Ai

Figure 3.2: Working of Bat Algorithm in Sentiment Analysis

```
For each bat $b_i$ (i = 1..m),
    For each feature j, do
        $X_i^j$ <- Random {0, 1}
    $v_i^j$ <- 0
    $A_i$ <- Random [1, 2]
    $r_i$ <-Random [0, 1]
    fit <- (-infinity)
global fit <- (-infinity)
For each iteration t (t = 1, …, T), do
    For each bat bi, do
        Create Z1 and Z2 from Z1 and Z2 respectively, such that both contains only features in bi in
        which xij is not equal to 0,
        Train Classifier over Z’1, evaluate its over Z’2 and stores the accuracy in acc
        rand <- random [0, 1]
        If (rand < $A_i$ and acc > fit), then
            fiti <- acc
            $A_i$ <- $\alpha A_i$
            $r_i \leftarrow r^0{}_i$ [1-erp (-$\gamma$t)]
        [maxfit, maxindex] ← max (fit)
        If (maxfit>globalfit), then
            globalfit ← maxfit
            For each dimension j (ɏj = 1…m), do
                $\ddot{x}j \leftarrow x^j{}_{maxindex}$
```

```
          For each bat bi (ɏi = 1….m), do
              β←Random [0,1]
              rand ← Random[0,1]
              If (rand > $r_i$)
                  For each feature j ( ɏj = 1…..n), do
                      $f_i \leftarrow f_{min} + (f_{max} – f_{min})\beta$
                      $v^j_i \leftarrow v^j_i + (\ddot{x}^j - x^j_i)fi$
                      $x^j_i \leftarrow x^j_i + v^j_i$
                      $\delta \leftarrow$ Random {0,1}
                      If ($\delta < \frac{1}{1+e-x}$), then
                          $x^j_i \leftarrow 1$
                      else $x^j_i \leftarrow 0$
For each feature j (ɏj = 1…n), do
    $F^j \leftarrow \ddot{x}^j$
Return F.
```

Figure 3.3: Bat Algorithm used in Feature Selection

3.3 Chapter Summary

This chapter explains the proposed Bat Algorithm for extracting Feature subset for classification in Sentiment Analysis. Also, it illustrates the working of Bat Algorithm in selecting Features.

CHAPTER 4

IMPLEMENTATION

In this chapter we will discuss the experimental setup of the research work done. First section will discuss the data set followed by feature vector and tools used for programming. In the next section, Nature-inspired optimization method is discussed. In the last section summary of the chapter is given.

4.1 Data Set

The data set used in this research is collected from social networking site – Twitter, i.e. tweets. The results are evaluated using two different size of dataset, dataset 1 is of small size and then the same algorithm is applied on large set of twitter and the results are compared.

4.2 Optimization Algorithm

Bat Algorithm and its improved version Binary Bat Algorithm is used as an optimization algorithm.

4.3 Programming Tool

Support Vector machine classifier is used to classify the features. Firstly, training classifier is done by using pre-classified training data and then its evaluation is checked. All the programming is done is Python programming language. Later the same dataset are compared using Weka tool without using Bat Algorithm.

4.4 Evaluation Methods

To evaluate the performance of the different classifiers, the accuracy of each separate classifier is computed. Accuracy measures the percentage of input in the test set that the classifier has labelled correctly. Furthermore, the precision and recall are calculated. Precision is the number of true positives divided by the total number of elements labelled as belonging to that class. A high precision means that the majority of items labelled as for instance ‘positive’ indeed belong to the class ‘positive’. Recall is the number of true positives divided by the total number of items that actually belong to that class. A high recall means that the majority of the ‘positive’ items were labelled as belonging to the class ‘positive’. The f-measure or f-score combines the precision and recall giving a single score, and is defined to be the harmonic mean of the precision and recall.

Description and outputs

Programming Tools and software used

Operating System	: Windows 10
Language used	: Python
Library used	: NLTK, Libsvm
Mining Tool	: Weka 3.6

Collecting Tweets

To have Tweets we need to have a twitter account on Dev twitter. Following steps are followed to create the account:

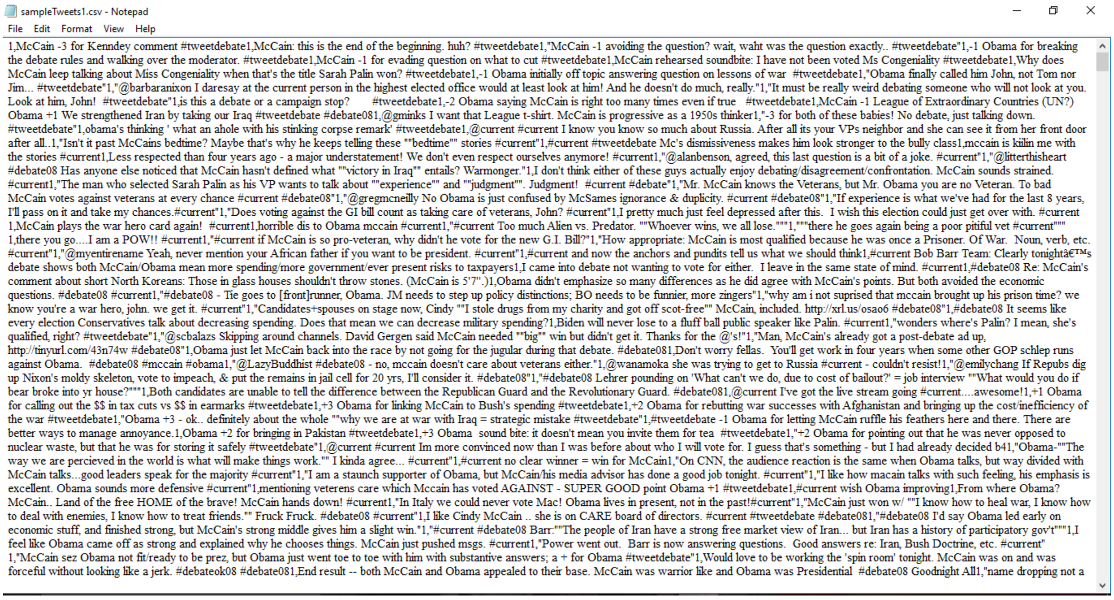
- Make a twitter account
- Open https://dev.twitter.com/ and login.
- Create a new application.
- Fill all the particulars and mention the redirect link.
- This will generate a consumer secret and consumer key, next access tokens are generated from access token tabs.
- We need to use these 4 keys to access Twitter data.

Figure 4.1 shows the screenshot of dev twitter app from where the four access keys are generated. The set of access keys used in this work are:

Keys	Key Value
Consumer Key (API Key)	DZL3nJm7HcwQDQveyhsMSAWnu
Consumer Secret (API Secret)	sbEM3uGIQQnDfKQGMufSiH07ECfCjd2yzceqqIvx1fruUawKcn
Access Token:	730629413214257152-zFphV7jZNKj0wiBv8CJaJfLacNlQgJq
Access Token Secret	tNieakSqbmWHbnbeZn9xIEznUWpYVBxNbsTPYPDrEtD8c

Table 4.1: Key value for Dev Twitter

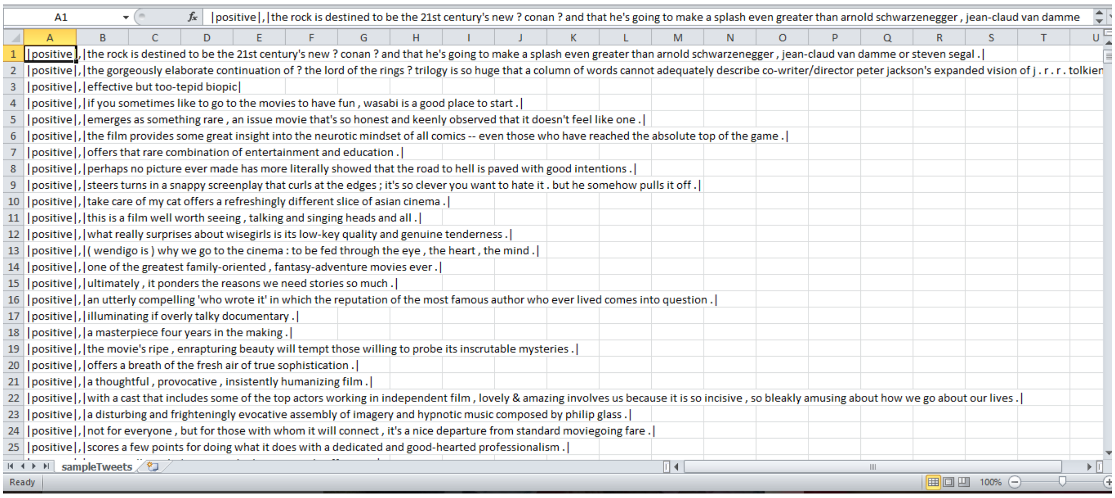
Tweets collected in notepad



```
1,McCain -3 for Kenndey comment #tweetdebate1,McCain: this is the end of the beginning. huh? #tweetdebate1,"McCain -1 avoiding the question? wait, waht was the question exactly.. #tweetdebate"1,-1 Obama for breaking
the debate rules and walking over the moderator. #tweetdebate1,McCain -1 for evading question on what to cut #tweetdebate1,McCain rehearsed soundbite: I have not been voted Ms Congeniality #tweetdebate1,Why does
McCain leep talking about Miss Congeniality when that's the title Sarah Palin won? #tweetdebate1,-1 Obama initially off topic answering question on lessons of war  #tweetdebate1,"Obama finally called him John, not Tom nor
Jim... #tweetdebate"1,"@barbaranixon I daresay at the current person in the highest elected office would at least look at him! And he doesn't do much, really."1,"It must be really weird debating someone who will not look at you.
Look at him, John!  #tweetdebate"1,is this a debate or a campaign stop?       #tweetdebate1,-2 Obama saying McCain is right too many times even if true   #tweetdebate1,McCain -1 League of Extraordinary Countries (UN?)
Obama +1 We strengthened Iran by taking our Iraq #tweetdebate #debate081,@gminks I want that League t-shirt. McCain is progressive as a 1950s thinker1,"-3 for both of these babies! No debate, just talking down.
#tweetdebate"1,obama's thinking ' what an ahole with his stinking corpse remark' #tweetdebate1,@current #current I know you know so much about Russia. After all its your VPs neighbor and she can see it from her front door
after all..1,"Isn't it past McCains bedtime? Maybe that's why he keeps telling these ""bedtime"" stories #current"1,#current #tweetdebate Mc's dismissiveness makes him look stronger to the bully class1,mccain is kiilin me with
the stories #current1,Less respected than four years ago - a major understatement! We don't even respect ourselves anymore! #current1,"@alanbenson, agreed, this last question is a bit of a joke. #current"1,"@litterthisheart
#debate08 Has anyone else noticed that McCain hasn't defined what ""victory in Iraq"" entails? Warmonger."1,I don't think either of these guys actually enjoy debating/disagreement/confrontation. McCain sounds strained.
#current1,"The man who selected Sarah Palin as his VP wants to talk about ""experience"" and ""judgment"". Judgment!  #current #debate"1,"Mr. McCain knows the Veterans, but Mr. Obama you are no Veteran. To bad
McCain votes against veterans at every chance #current #debate08"1,"@gregmcneilly No Obama is just confused by McSames ignorance & duplicity. #current #debate08"1,"If experience is what we've had for the last 8 years,
I'll pass on it and take my chances.#current"1,"Does voting against the GI bill count as taking care of veterans, John? #current"1,I pretty much just feel depressed after this.  I wish this election could just get over with. #current
1,McCain plays the war hero card again!  #current1,horrible dis to Obama mccain #current1,"#current Too much Alien vs. Predator. ""Whoever wins, we all lose."""1,"""there he goes again being a poor pitiful vet #current"""
1,there you go....I am a POW!! #current1,"#current if McCain is so pro-veteran, why didn't he vote for the new G.I. Bill?"1,"How appropriate: McCain is most qualified because he was once a Prisoner. Of War.  Noun, verb, etc.
#current"1,"@myentirename Yeah, never mention your African father if you want to be president. #current"1,#current and now the anchors and pundits tell us what we should think1,#current Bob Barr Team: Clearly tonightâ€™s
debate shows both McCain/Obama mean more spending/more government/ever present risks to taxpayers1,I came into debate not wanting to vote for either.  I leave in the same state of mind. #current1,#debate08 Re: McCain's
comment about short North Koreans: Those in glass houses shouldn't throw stones. (McCain is 5'7''.)1,Obama didn't emphasize so many differences as he did agree with McCain's points. But both avoided the economic
questions. #debate08 #current1,"#debate08 - Tie goes to [front]runner, Obama. JM needs to step up policy distinctions; BO needs to be funnier, more zingers"1,"why am i not suprised that mccain brought up his prison time? we
know you're a war hero, john. we get it. #current"1,"Candidates+spouses on stage now, Cindy ""I stole drugs from my charity and got off scot-free"" McCain, included. http://xrl.us/osao6 #debate08"1,#debate08 It seems like
every election Conservatives talk about decreasing spending. Does that mean we can decrease military spending?1,Biden will never lose to a fluff ball public speaker like Palin. #current1,"wonders where's Palin? I mean, she's
qualified, right? #tweetdebate"1,"@scbalazs Skipping around channels. David Gergen said McCain needed ""big"" win but didn't get it. Thanks for the @'s!"1,"Man, McCain's already got a post-debate ad up,
http://tinyurl.com/43n74w #debate08"1,Obama just let McCain back into the race by not going for the jugular during that debate. #debate081,Don't worry fellas.  You'll get work in four years when some other GOP schlep runs
against Obama.  #debate08 #mccain #obama1,"@LazyBuddhist #debate08 - no, mccain doesn't care about veterans either."1,@wanamoka she was trying to get to Russia #current - couldn't resist!1,"@emilychang If Repubs dig
up Nixon's moldy skeleton, vote to impeach, & put the remains in jail cell for 20 yrs, I'll consider it. #debate08"1,"#debate08 Lehrer pounding on 'What can't we do, due to cost of bailout?' = job interview ""What would you do if
bear broke into yr house?"""1,Both candidates are unable to tell the difference between the Republican Guard and the Revolutionary Guard. #debate081,@current I've got the live stream going #current....awesome!1,+1 Obama
for calling out the $$ in tax cuts vs $$ in earmarks #tweetdebate1,+3 Obama for linking McCain to Bush's spending #tweetdebate1,+2 Obama for rebutting war successes with Afghanistan and bringing up the cost/inefficiency of
the war #tweetdebate1,"Obama +3 - ok.. definitely about the whole ""why we are at war with Iraq = strategic mistake #tweetdebate"1,#tweetdebate -1 Obama for letting McCain ruffle his feathers here and there. There are
better ways to manage annoyance.1,Obama +2 for bringing in Pakistan #tweetdebate1,+3 Obama  sound bite: it doesn't mean you invite them for tea  #tweetdebate1,"+2 Obama for pointing out that he was never opposed to
nuclear waste, but that he was for storing it safely #tweetdebate"1,@current #current Im more convinced now than I was before about who I will vote for. I guess that's something - but I had already decided b41,"Obama-""The
way we are percieved in the world is what will make things work."" I kinda agree... #current"1,#current no clear winner = win for McCain1,"On CNN, the audience reaction is the same when Obama talks, but way divided with
McCain talks...good leaders speak for the majority #current"1,"I am a staunch supporter of Obama, but McCain/his media advisor has done a good job tonight. #current"1,"I like how macain talks with such feeling, his emphasis is
excellent. Obama sounds more defensive #current"1,mentioning veterens care which Mccain has voted AGAINST - SUPER GOOD point Obama +1 #tweetdebate1,#current wish Obama improving1,From where Obama?
McCain.. Land of the free HOME of the brave! McCain hands down! #current1,"In Italy we could never vote Mac! Obama lives in present, not in the past!#current"1,"McCain just won w/ ""I know how to heal war, I know how
to deal with enemies, I know how to treat friends."" Fruck Fruck. #debate08 #current"1,I like Cindy McCain .. she is on CARE board of directors. #current #tweetdebate #debate081,"#debate08 I'd say Obama led early on
economic stuff, and finished strong, but McCain's strong middle gives him a slight win."1,"#current #debate08 Barr:""The people of Iran have a strong free market view of Iran... but Iran has a history of participatory gov't"""1,I
feel like Obama came off as strong and explained why he chooses things. McCain just pushed msgs. #current1,"Power went out.  Barr is now answering questions.  Good answers re: Iran, Bush Doctrine, etc. #current"
1,"McCain sez Obama not fit/ready to be prez, but Obama just went toe to toe with him with substantive answers; a + for Obama #tweetdebate"1,Would love to be working the 'spin room' tonight. McCain was on and was
forceful without looking like a jerk. #debateok08 #debate081,End result -- both McCain and Obama appealed to their base. McCain was warrior like and Obama was Presidential  #debate08 Goodnight All1,"name dropping not a
```

Figure 4.1: Tweets collected

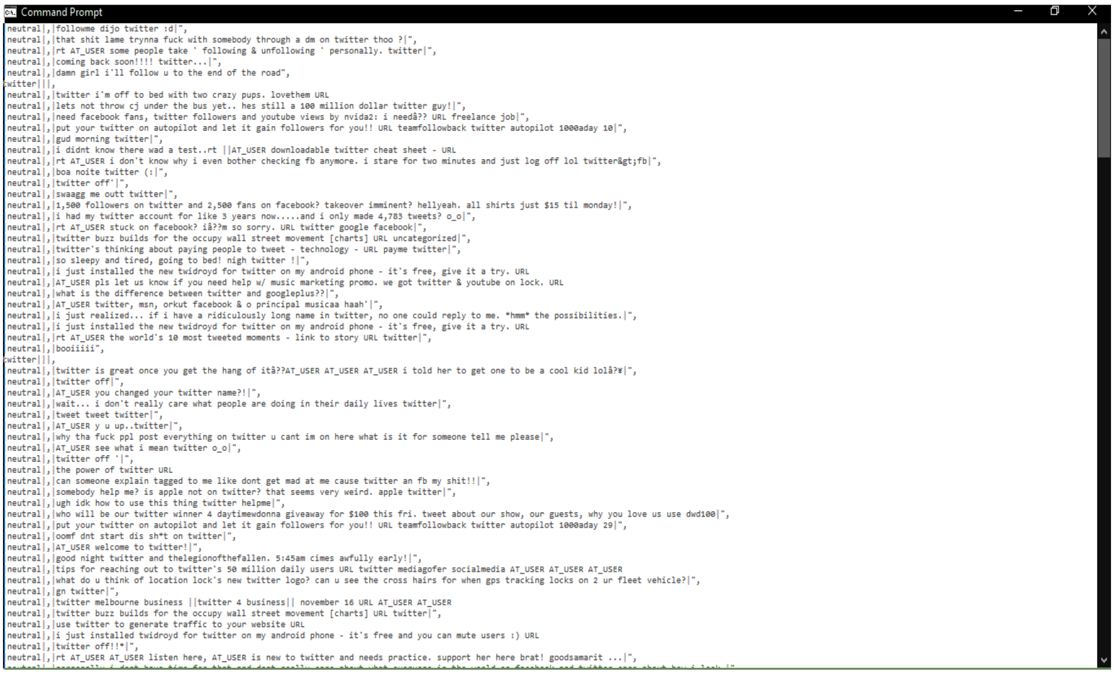
Tweets collected in .csv format



```
A1    |positive|,|the rock is destined to be the 21st century's new ? conan ? and that he's going to make a splash even greater than arnold schwarzenegger , jean-claud van damme
1  |positive|,|the rock is destined to be the 21st century's new ? conan ? and that he's going to make a splash even greater than arnold schwarzenegger , jean-claud van damme or steven segal .|
2  |positive|,|the gorgeously elaborate continuation of ? the lord of the rings ? trilogy is so huge that a column of words cannot adequately describe co-writer/director peter jackson's expanded vision of j . r . r . tolkien
3  |positive|,|effective but too-tepid biopic|
4  |positive|,|if you sometimes like to go to the movies to have fun , wasabi is a good place to start .|
5  |positive|,|emerges as something rare , an issue movie that's so honest and keenly observed that it doesn't feel like one .|
6  |positive|,|the film provides some great insight into the neurotic mindset of all comics -- even those who have reached the absolute top of the game .|
7  |positive|,|offers that rare combination of entertainment and education .|
8  |positive|,|perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions .|
9  |positive|,|steers turns in a snappy screenplay that curls at the edges ; it's so clever you want to hate it . but he somehow pulls it off .|
10 |positive|,|take care of my cat offers a refreshingly different slice of asian cinema .|
11 |positive|,|this is a film well worth seeing , talking and singing heads and all .|
12 |positive|,|what really surprises about wisegirls is its low-key quality and genuine tenderness .|
13 |positive|,|( wendigo is ) why we go to the cinema : to be fed through the eye , the heart , the mind .|
14 |positive|,|one of the greatest family-oriented , fantasy-adventure movies ever .|
15 |positive|,|ultimately , it ponders the reasons we need stories so much .|
16 |positive|,|an utterly compelling 'who wrote it' in which the reputation of the most famous author who ever lived comes into question .|
17 |positive|,|illuminating if overly talky documentary .|
18 |positive|,|a masterpiece four years in the making .|
19 |positive|,|the movie's ripe , enrapturing beauty will tempt those willing to probe its inscrutable mysteries .|
20 |positive|,|offers a breath of the fresh air of true sophistication .|
21 |positive|,|a thoughtful , provocative , insistently humanizing film .|
22 |positive|,|with a cast that includes some of the top actors working in independent film , lovely & amazing involves us because it is so incisive , so bleakly amusing about how we go about our lives .|
23 |positive|,|a disturbing and frighteningly evocative assembly of imagery and hypnotic music composed by philip glass .|
24 |positive|,|not for everyone , but for those with whom it will connect , it's a nice departure from standard moviegoing fare .|
25 |positive|,|scores a few points for doing what it does with a dedicated and good-hearted professionalism .|
```

Figure 4.2 Tweets in .csv format

Preprocessing of tweets



```
neutral|,|followme dijo twitter :d|",
neutral|,|that shit lame trynna fuck with somebody through a dm on twitter thoo ?|",
neutral|,|rt AT_USER some people take ` following & unfollowing ` personally. twitter|",
neutral|,|coming back soon!!!! twitter...|",
neutral|,|damn girl i'll follow u to the end of the road",
twitter|||,
neutral|,|twitter i'm off to bed with two crazy pups. lovethem URL
neutral|,|lets not throw cj under the bus yet.. hes still a 100 million dollar twitter guy!|",
neutral|,|need facebook fans, twitter followers and youtube views by nvida2: i needâ?? URL freelance job|",
neutral|,|put your twitter on autopilot and let it gain followers for you!! URL teamfollowback twitter autopilot 1000aday 10|",
neutral|,|gud morning twitter|",
neutral|,|i didnt know there wad a test..rt ||AT_USER downloadable twitter cheat sheet - URL
neutral|,|rt AT_USER i don't know why i even bother checking fb anymore. i stare for two minutes and just log off lol twitter&gt;fb|",
neutral|,|boa noite twitter (:|",
neutral|,|twitter off'|",
neutral|,|swaagg me outt twitter|",
neutral|,|1,500 followers on twitter and 2,500 fans on facebook? takeover imminent? hellyeah. all shirts just $15 til monday!|",
neutral|,|i had my twitter account for like 3 years now.....and i only made 4,783 tweets? o_o|",
neutral|,|rt AT_USER stuck on facebook? iâ??m so sorry. URL twitter google facebook|",
neutral|,|twitter buzz builds for the occupy wall street movement [charts] URL uncategorized|",
neutral|,|twitter's thinking about paying people to tweet - technology - URL payme twitter|",
neutral|,|so sleepy and tired, going to bed! nigh twitter !|",
neutral|,|i just installed the new twidroyd for twitter on my android phone - it's free, give it a try. URL
neutral|,|AT_USER pls let us know if you need help w/ music marketing promo. we got twitter & youtube on lock. URL
neutral|,|what is the difference between twitter and googleplus??|",
neutral|,|AT_USER twitter, msn, orkut facebook & o principal musicaa haah'|",
neutral|,|i just realized... if i have a ridiculously long name in twitter, no one could reply to me. *hmm* the possibilities.|",
neutral|,|i just installed the new twidroyd for twitter on my android phone - it's free, give it a try. URL
neutral|,|rt AT_USER the world's 10 most tweeted moments - link to story URL twitter|",
neutral|,|booiiiii",
twitter|||,
neutral|,|twitter is great once you get the hang of itâ??AT_USER AT_USER AT_USER i told her to get one to be a cool kid lolâ?¥|",
neutral|,|twitter off|",
neutral|,|AT_USER you changed your twitter name?!|",
neutral|,|wait... i don't really care what people are doing in their daily lives twitter|",
neutral|,|tweet tweet twitter|",
neutral|,|AT_USER y u up..twitter|",
neutral|,|why tha fuck ppl post everything on twitter u cant im on here what is it for someone tell me please|",
neutral|,|AT_USER see what i mean twitter o_o|",
neutral|,|twitter off '|",
neutral|,|the power of twitter URL
neutral|,|can someone explain tagged to me like dont get mad at me cause twitter an fb my shit!!|",
neutral|,|somebody help me? is apple not on twitter? that seems very weird. apple twitter|",
neutral|,|ugh idk how to use this thing twitter helpme|",
neutral|,|who will be our twitter winner 4 daytimewdonna giveaway for $100 this fri. tweet about our show, our guests, why you love us use dwd100|",
neutral|,|put your twitter on autopilot and let it gain followers for you!! URL teamfollowback twitter autopilot 1000aday 29|",
neutral|,|oomf dnt start dis sh*t on twitter|",
neutral|,|AT_USER welcome to twitter!|",
neutral|,|good night twitter and thelegionofthefallen. 5:45am cimes awfully early!|",
neutral|,|tips for reaching out to twitter's 50 million daily users URL twitter mediagofer socialmedia AT_USER AT_USER AT_USER
neutral|,|what do u think of location lock's new twitter logo? can u see the cross hairs for when gps tracking locks on 2 ur fleet vehicle?|",
neutral|,|gn twitter|",
neutral|,|twitter melbourne business ||twitter 4 business|| november 16 URL AT_USER AT_USER
neutral|,|twitter buzz builds for the occupy wall street movement [charts] URL twitter|",
neutral|,|use twitter to generate traffic to your website URL
neutral|,|i just installed twidroyd for twitter on my android phone - it's free and you can mute users :) URL
neutral|,|twitter off!!*|",
neutral|,|rt AT_USER AT_USER listen here, AT_USER is new to twitter and needs practice. support her here brat! goodsamarit ...|",
```

Figure 4.3: Prepressing of Tweets.

CHAPTER 5

RESULTS & ANALYSIS

The above codes are executed and results obtained are as follows:

Results on small corpus

Model	**Accuracy**
SVM (using weka)	74.20
SVM + Bat Algorithm	82.79
SVM + Binary Bat Algorithm	84.23

Table 5.1: Comparisons of results on small corpus

Results on large corpus

Model	**Accuracy**
SVM (using weka)	76.34
SVM + Bat Algorithm	81.23
SVM + Binary Bat Algorithm	83.92

Table 5.2: Comparisons of results on large corpus

As it can be observed following sets of tweets give better accuracy using Bat algorithm and SVM with that of only using SVM.

Also, Binary Bat algorithm shows that changing '0' and '1' on hyperplane gives more accurate results than Bat algorithm.

CHAPTER 6

CONCLUSION AND FUTURE WORKS

This chapter concludes the contributions made by this thesis. Also figure out the limitation of the work done and briefly discuss the future scope of the research.

6.1 Research Summary

The research in this work introduces an optimization algorithm for an optimal feature subset to achieve a remarkable accuracy. While detecting sarcasm, most methods fail to detect the minor negative words in a sentence which affects the polarity. So to achieve accurate results Nature Inspired Algorithms are used for optimizing algorithms. In this work we have discussed a novel method so that one doesn’t have to compromise with the optimal solution. Here we used Bat Algorithm which classifies the mood of the opinion holder. While selecting feature due to redundancy, classifiers often miss out some important feature which alter the meaning of the sentence due to which size of the corpus also increases proportionally which later on is difficult to maintain. Every so often, there is a probability of repeating features in the subset. According to the results obtain with Bat algorithm and without Bat algorithm shows a drastic change in feature selection subsets which leads to improvement of accuracy. Sentiment Analysis score is calculated in three polarity (Positive, Negative and Neutral).

6.2 Conclusion

In this research a novel method for feature selection has been evaluated. The main input in this work is to propose a method to select feature set that could lead to give almost correct results using training dataset. To increase classification accuracy we have optimized feature selection for reducing feature subset corpus size and computational complexity. Bat Algorithm and Binary Bat Algorithm are used with SVM to perceptively select the most promising features from search space which could maximize the accuracy classified by the classifier. The Redundancy and noisy features are deleted from feature set and are computed. From the study we conclude that Bat algorithm and Binary Bat Algorithm gives better results.

6.3 Future Scope

This study is empirical in nature and can be evaluated further using various Nature-inspired algorithms discussed in table 2.1. The results demonstrate that the discussed algorithm has the characteristics of distinguishing opinions. The algorithm has only been investigated on Twitter Dataset. The study can be carried further on different area of dataset using much larger samples. There are few assumptions, known limitations with respect to the analysis and data collected, also some technical challenges which may affect the accuracy and polarity of the results.

- For using Bat Algorithm, one must need to use a transfer function for solving high dimensionality and finding finest features in the given search space by representing it as a hypercube. Instead of changing Bat's position into binary representation we can assume a hypercube which changes its dimensions value bit by bit.

- Tweets collected are assumed to be in English language but this may not be possible every time. Users mostly share their opinions in the language they are most comfortable with and sometimes they use English with their inherent language.
  E.g. "Super-Awesome mausam:*:*"
  The word "mausam" in the above tweet will be difficult to interpret by the machine and will give a conflict to combine it.

References

1. Bo Pang., Lilliam Lee.: Opinion Mining and Sentiment Analysis. *Foundations* and Trends in Information Retrieval. Vol. 2, No 1-2 (2008) 1–13 (2008).

2. Akshi Kumar, Teeja Mary Sebastian, Sentiment Analysis: A Perspective on its Past, Present and Future, International Journal of Intelligent Systems and Applications, Vol.4, No.10, 2012.

3. Yiming Yang, Jan O. Pederson, A Comparative study on Feature Selection in Text Categorization (1997).

4. Ahmed Abbasi, et al. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums" ACM Transactions on Information Systems, Vol. 26, No. 3, 2008.

5. Mehdi Hosseinzadeh Aghdam *, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri, Text feature selection using ant colonyoptimization, Expert Systems with Applications 36 (2009) 6843–6853

6. A. Abbasi, et al.;Selecting Attributes for Sentiment Classification Using Feature Relation Networks,&quot; IEEE Transactions on Knowledge & Data Engineering, vol. 23, 447-462, 2011.

7. Ekbal, A., Saha, S., and Garbe, C. S. "Feature selection using multi objective optimization for named entity recognition" In proceedings of IEEE 20th International Conference on Pattern Recognition, pp. 1937-1940,2010.

8. Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In Proceedings of the18th international conference on World wide web, pages 131–140, Madrid, Spain, 2009, ACM.

9. Sangita Roy, Samir Biswas, Sheli Sinha Chaudhuri, Nature-Inspired Swarm Intelligence and Its Applications, I.J. Modern Education and Computer Science, 2014, 12, 55-65

10. Carlos M. Fonseca, Peter J. Fleming. An Overview of Evolutionary Algorithms in Multiobjective Optimization. Spring 1995, Vol. 3, No. 1, Pages 1-16 Massachusetts Institute of Technology, Online December 10, 2007.

11. Azizah Binti Mohamad, Azlan Mohd Zain & Nor Erne Nazira Bazin, Cuckoo Search Algorithm for Optimization Problems—A Literature Review and its Applications, Applied Artificial Intelligence, ISSN: 0883-9514.

12. O'Keefe, T., and Koprinska, I. "Feature selection and weighting methods in sentiment analysis", InProceedings of 14th Australasian Document Computing Symposium, pp-67- 74, 2009.

13. Rui Tang, S. Fong, Xin-She Yang, and S. Deb. Wolf search algorithm with ephemeral memory. In Digital Information Management (ICDIM), Seventh International Conference, 165–172, 2012.

14. Amir Hossein Gandomi and Amir Hossein Alavi. Krill herd: a new bio-inspired optimization algorithm. Communications in Nonlinear Science and Numerical Simulation, 2012.

15. William L. Goffe, Gary D. Ferrier, John Rogers, Global optimization of statistical functions with simulated annealing, Journal of EconometricsVolume 60, Issues 1–2, 1994, pp 65-99

16. Kevin M Passino. Biomimicry of bacterial foraging for distributed optimization and control. Control Systems, IEEE, 22(3):52–67, 2002.

17. Xin-She Yang, Janet M Lees, and Chris T Morley. Application of virtual ant algorithms in the optimization of cfrp shear strengthened precracked structures. In Computational Science– ICCS 2006, pages 834–837. Springer, 2006.

18. X.-L. Li, Z.-J. Shao, and J.-X. Qian. Optimizing method based on autonomous animats: Fish-swarm algorithm. Xitong Gongcheng Lilunyu Shijian/System Engineering Theory and Practice, 22(11):32, 2002.

19. Ruby Dhurve, Megha Seth, " Weighted Sentiment Analysis Using Artificial Bee Colony Algorithm", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 James Kennedy and Russell Eberhart. Particle swarm optimization. InNeural Networks, 1995.

20. James Kennedy and Russell Eberhart. Particle swarm optimization. InNeural Networks, 1995. Proceedings., IEEE International Conference on, volume 4, pages 942–948. IEEE, 1995.

21. George Stylios, Christos D. Katsis, DimitrisChristodoulakis, " Using Bio-inspired Intelligence for Web Opinion Mining", International Journal of Computer Applications Vol 87 – No.5, 2014.

22. DT Pham, A Ghanbarzadeh, E Koc, S Otri, S Rahim, and M Zaidi. The bees algorithm-a novel tool for complex optimisation problems. Proceedings of the 2nd Virtual International Conference on Intelligent Production Machines and Systems, pages 454–459, 2006.

23. T. Sumathi, S.Karthik, M.Marikkannan, "Artificial Bee Colony Optimization for Feature Selection in Opinion Mining", Journal of Theoretical and Applied Information Technology, 10th august 2014. vol. 66 no.1.

24. H.F. Wedde, M. Farooq, and Y. Zhang. Beehive: An efficient fault-tolerant routing algorithm inspired by honey bee behavior. Lecture Notes in Computer Science 3172 LNCS:83–94, 2004.

25. M. Dorigo, Optimization, Learning and Natural Algorithms, PhD thesis, Politecnico di Milano, Italy, 1992.

26. P Lucic and D Teodorovic. Bee system: modeling combinatorial optimization transportation engineering problems by swarm intelligence. In Preprints of the TRISTAN IV triennial symposium on transportation analysis, pages 441–445, 2001.

27. Abd. Samad Hasan Basari, BurairahHussin, I. GedePramudya Ananta, Junta Zeniarja, "Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization".

28. Dervis Karaboga, An Idea Based On Honey Bee Swarm for Numerical Optimization, Technical Report-TR06, Erciyes University,Engineering Faculty, Computer Engineering Department 2005.

29. Dusˇan Teodorovic´ and Mauro Dell'Orco. Bee colony optimization–a cooperative learning approach to complex transportation problems. In Advanced OR and AI Methods in Transportation: Proceedings of 16th Mini–EURO Conference and 10th Meeting of EWGT 2005.

30. Habiba Drias, Souhila Sadeg, and Safa Yahi. Cooperative bees swarm for solving the maximum weighted satisfiability problem. In Computational Intelligence and Bioinspired Systems, pages 318–325. Springer, 2005.

31. X.-S. Yang. Engineering optimizations via nature-inspired virtual bee algorithms. volume 3562, pages 317–323, 2005.

32. Deepak Kumar Gupta, Kandula Srikanth Reddy, Shweta, Asif Ekbal, "PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis", Natural Language Processing and Information Systems,Volume 9103 of the series Lecture Notes in Computer Science pp 220-233

33. Shoubao Su, Jiwen Wang, Wangkang Fan, and Xibing Yin. Good lattice swarm algorithm for constrained engineering design optimization. In Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. pages 6421–6424. IEEE, 2007.

34. KN Krishnanand and D Ghose. Detection of multiple source locations using a glowworm metaphor with applications to collective robotics. In Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE, pages 84–91. IEEE, 2005.

35. Antonio Mucherino and Onur Seref. Monkey search: a novel metaheuristic search for global optimization. In Data Mining, Systems Analysis and Optimization in Biomedicine, volume 953, pages 162–173, 2007.

36. Ying Chu, Hua Mi, Huilian Liao, Zhen Ji, and QH Wu. A fast bacterial swarming algorithm for high-dimensional function optimization. In Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence), pages 3135–3140. IEEE, 2008.

37. Xin-She Yang. Firefly algorithm, stochastic test functions and design optimization. International Journal of Bio-Inspired Computation, 2(2):78–84, 2010.

38. X.S. Yang. A new metaheuristic bat-inspired algorithm. Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pages 65–74, 2010.

39. Serban Iordache. Consultant-guided search: a new metaheuristic for combinatorial optimization problems. In Proceedings of the 12th annual conference on Genetic and evolutionary computation, pages 225–232. ACM, 2010.

40. Xin-She Yang and Suash Deb. Eagle strategy using le´vy walk and firefly algorithms for stochastic optimization. In Nature Inspired Cooperative Strategies for Optimization (NICSO2010), pages 101–111.Springer, 2010.

41. TO Ting, Ka Lok Man, Sheng-Uei Guan, Mohamed Nayel, and Kaiyu Wan. Weightless swarm algorithm (wsa) for dynamic optimization problems. In Network and Parallel Computing, pages 508–515. Springer, 2012.

42. Francesc de Paula Comellas Padro´, Jesu´s Mart´ınez Navarro, et al. Bumblebees: a multiagent combinatorial optimization algorithm inspired by social insect behaviour. 2011.

43. JURAFSKY, D. & MARTIN, J.H. (2009), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Pearson Education, Inc

44. JURAFSKY, D. & MARTIN, J.H. (2009), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. New Jersey: Pearson Education, Inc

45. FELDMAN, R. & SANGER, J. (2007), The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press.]

46. Rui Xia , Chengqing Zong, Shoushan Li, “Ensemble of feature sets and classification algorithms for sentiment classification”, Information Sciences 181 (2011) 1138–1152.

47. FRADKIN, D. & MUCHNIK, I. (2006), “Support Vector Machines for Classification”. In ABELLO, J. & CARMODE, G. [Eds.], Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 70, p. 13-20

48. ref-RUSSELL, S. & NORVIG, P. (2003), Artificial Intelligence: A Modern Approach. New Jersey: Pearson Education, Inc.]

49. NG, V. & CARDIE, C. (2003), “Weakly-Supervised Natural Language Learning Without Redundant Views”. In Proceedings of the Conference on Human Language Technologies – North American Chapter of the Association for Computational Linguistics (HLT-NAACL-2003), p. 94-101.

50. George H John, Ron Kohavi, Karl Pfleger, et al. “Irrelevant Features and the Subset Selection Problem.” In: ICML. Vol. 94. 1994, pp. 121–129

51. S. Mirjalili, S. M. Mirjalili, X. Yang, Binary Bat Algorithm, Neural Computing and Applications, In press, 2014, Springer DOI: http://dx.doi.org/10.1007/s00521-013-1525-5

52. Nakamura, Rodrigo YM, et al. "BBA: a binary bat algorithm for feature selection." *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2012.

53. X. S. Yang and A. H. Gandomi, Bat algorithm: a novel approach for global engineering optimization, Engineering Computations, Vol. 29, No. 5, pp. 464-483 (2012).].

54. K. Khan and A. Sahai, A comparison of BA, GA, PSO, BP and LM for training feed forward neural networks in e-learning context, Int. J. Intelligent Systems and Applications (IJISA), Vol. 4, No. 7, pp. 23-29 (2012).

55. Yang, Xin-She. "A new metaheuristic bat-inspired algorithm." *Nature inspired cooperative strategies for optimization (NICSO 2010)*. Springer Berlin Heidelberg, 2010. 65-74.

56. Sabba, Sara, and Salim Chikhi. "A discrete binary version of bat algorithm for multidimensional knapsack problem." *International Journal of Bio-Inspired Computation* 6.2 (2014): 140-152.