Hybrid Intelligent Techniques in Text Mining and Analysis of Social Networks and Media Data

Neha Golani, Ishan Khandelwal, and B.K. Tripathy

Abstract Text data from social media and networks are ubiquitous and are emerging at a high rate. Tackling these bulky text data has become a challenging task and an important field of research. The mining of text data and examining it with the help of several clustering techniques, classification techniques, and soft computing methods has been studied in a comprehensive manner in the past. This chapter focuses mainly on the hybrid techniques that have been used to mine textual data from social networks and media data. Social networks are considered a profuse source of viewpoints and outlooks of the public on a worldwide scale. An enormous amount of social media data is produced on a regular basis, generated because of the communication between the users who have signed up for the various social media platforms on several topics such as books, movies, politics, products, etc. The users vary in terms of factors such as viewpoints, scenarios, geographical situations, and many other settings. If mined efficiently, the data have the potential to provide a helpful outcome of an exegesis of social quirks and traits. This chapter offers a detailed methodology on how data mining, especially text mining, is applied to social networks in general. Furthermore, it goes on to introduce the traditional models used in mining the various hybrid methodologies that have evolved and make a comparative analysis. We also aim to provide the future scope and research studies present in this field with all possible new innovations and their applications in the real world.

Keywords Ant colony optimisation • Hybrid techniques • Neural networks • Particle swarm optimisation • Social networks • Support vector machine • Text mining

1 Introduction

A vast amount of research has been done with respect to techniques in Text Mining, Analysis of Social Networks and Media Data. An application has been created that intends to take the information in text form linked to the data that is pertinent to

N. Golani (✉) • I. Khandelwal • B.K. Tripathy
VIT University, Vellore, India
e-mail: nehagolani00@gmail.com; ishankhandelwal23@gmail.com; tripathybk@vit.ac.in

H. Banati et al. (eds.), *Hybrid Intelligence for Social Networks*,
DOI 10.1007/978-3-319-65139-2_1

identifying the geographical location of a person or device by means of digital information processed via the Internet. These data are taken as input and data mining is performed in the settings where the network of the social media data is complex. The idea is to extract the information, which works on social recommendation. It has addressed the challenge of the inability of the recommender systems to refine the search and posit the suggestions to the application users. It also puts forth the concept of Social Network based on Location linked with methodologies of text mining, and discloses issues that still require research for more efficacious and integrated outcomes.

The opinion mining techniques are applied to social media data such as Twitter data, which proves to be an effective means of portraying the organisations in terms of speed and efficiency of delivering the message along with the wide range of audience covered. Several features and techniques for training the opinion classifiers for Twitter datasets have been studied in the past few years, with different results. The problems of the previous conventional techniques are the precision of classification, the data being sparse, and sarcasm, as they incorrectly categorise the Tweets in the opposite category or neutral owing to the failure to understand the previously mentioned factors.

Social media have had an extremely significant impact on our lives and on our interactions with others. This makes it necessary for a successfully working organisation to have the capability to analyse the current occurrences using the information available, which involves reviews and experiences of customers of the company's services and products to predict the near future. This would assist the company to develop a better customer experience and ultimately build a better stature for the company. Several companies do not have enough knowledge on the data mining of social media efficaciously and they are not acquainted with the competitive intelligence of social media. Because of the abilities of text mining, it can be concluded that the implementation of text data to social media data can provide fascinating insights into human interactions and psychology.

Moreover, one of the endowments of the research relates to the study of potential linguistic tags as they enable the service-users to develop a vocabulary of their own and eventually examine the investigated or mapped areas and impart knowledge on these. They aspire to extend this technique with the assistance of similarity metrics to build personalised recommendations on the Web using collaborative filtration of data in Data Mining, culminating in a hybrid outlook.

Soft computing (SC) constitutes various paradigms of computing, which involve neural networks (NNs), genetic algorithms (GAs), fuzzy logic (FL), which are helpful in generating substantial and effective hybrid intelligent systems for text mining. An immense amount of research has been done in the field of hybrid methods of text mining. Kumar et al. [13] proposed a neuro-fuzzy method for mining of social media data with the help of the Cloud. Figure 1 [22] depicts the basic methodology for the analysis and mining of big text data.

The process of text mining of the social media data can be categorised into three steps. The first step is the text pre-processing step in which the textual data are extracted from the social media sites, collected and stored as the preparation of text

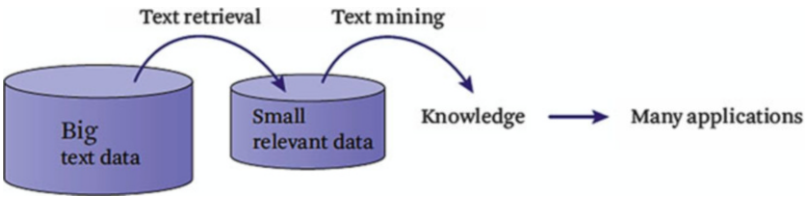

Fig. 1 Methodology for analysing big text data: text retrieval and text mining [22]

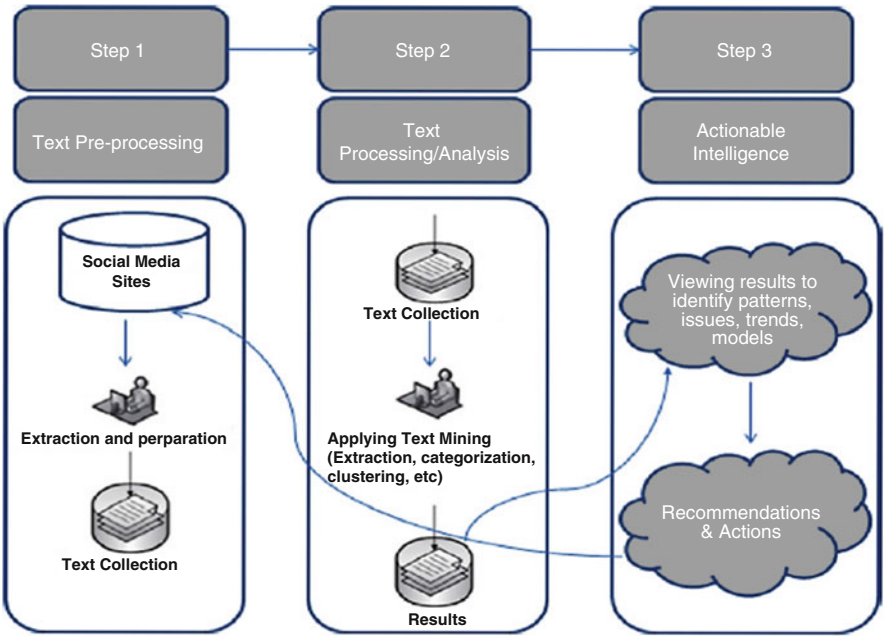

Fig. 2 Process of text mining for social media content [6]

processing and analysis. In the text processing/analysis step, an assortment of text mining techniques are applied to the data collected in the first step. The final step is the actionable intelligence step in which the results obtained from the previous steps are further processed to identify the various patterns, issues, trends, and models and to provide vital recommendations and actions based on the knowledge obtained (see Fig. 2).

The social network data are not very unusual from the conventional data, although a specialised language used for social media data to explain the skeleton and indices of the sets of observations that are usually taken into consideration. Furthermore, the datasets, which are developed by the networkers, generally result in different looking data array from the traditional rectangular data array with which the researchers and analysts are acquainted. One of the important differences between conventional data and network data is that with conventional data factors

such as actors and attributes are considered, whereas with network data more emphasis is placed on the actors and relations. This creates a difference in approach that a researcher has to take in terms of generation of a research design.

The chapter has begun with a brief introduction on social network and media data. Next, a concise introduction is given on text mining, sentiment analysis, and opinion mining. The next section focuses on all the traditional methods being used and introduces an in-depth analysis of the mathematics required [1, 21]. Then, the chapter moves on to the hybrid techniques being used in social media and how useful they are. The chapter includes hybrid intelligence used in Text clustering, text categorisation, opinion mining and sentiment analysis. We then go on to make a comparative analysis on how different hybrid techniques are more efficient and give better results than the traditional concepts. We finally end with the section that describes the scope of future research and an extensive bibliography.

2 Literature Review

The purpose of the chapter [5] was to present a methodology for performing the Social Recommendation-focused approach to filtering data based on content. The work also includes a study on the structure of data entities in the Social Network (structured in the Java Script Object Notation format) and designing a crawler to collect data that can still be used in other approaches needing to extract data from a social network. One of the contributions of the research concerns the study of potential semantic tags because it allows the users to create their own vocabulary and spread knowledge discovery in unexplored areas. The proposed algorithm is used to search through the scores and the cumulative sums approaching the similarity between items, or other profiles sought at the time of the user's query. The technique seeks to reduce the ambiguities and redundancies that are found in terms of semantic relations. The main contribution of this work is undoubtedly the creation of a new methodology that could be adopted in the recommendation process on location based social networks web environment, where further research is still required in addition to consultation with the user and the use of data mining techniques in text, with consolidated results.

In [13], the authors have tried to analyse the advantages and pitfalls of artificial neural networks (ANN) and fuzzy approaches to mining social media datasets. They analysed Web mining and its types such as classification and clustering. It gives us an insight into Web usage mining through artificial neural networks, the use of fuzzy logic and ant colony optimisation (ACO) in web mining. It suggests the use of Social Network Analysis (SNA) as an essential tool for researchers owing to the increase in the number of social media users. It enlists merits and demerits of several methods in soft computing, such as genetic algorithm, artificial neural network, ant colony optimisation and fuzzy set for mining the datasets of social media.

In [10], the authors have implemented and evaluated Naïve Bayes and J48 algorithms under two-term frequency and TF-IDF (Term Frequency Inverse

Document Frequency) term weighting methods. They concluded that the J48 method performs better than the Naïve Bayes in terms of frequency and TF-IDF term weighting methods. It is concluded that in text classification, the J48 method outperforms Naïve Bayes in term frequency and TF-IDF term weighting methods. Also, according to results obtained, TF-IDF has better performance than the term frequency method in text classification.

Furthermore in [15], to categorise each opinion given by the viewers as positive or negative for a particular review dataset, hybrid methods are put forward for consideration. The reviews and ratings of movies on Twitter are considered in Twitter with the help of sentiment analysis. A hybrid method using particle swarm optimisation (PSO) and support vector machine (SVM) is used to categorise the opinions of the user into positive, negative, for a remarks dataset of a particular movie. These results are helpful in an improved decision-making process.

In [12], the authors have proposed a novel hybrid approach for determining the sentiment of individual Tweets on Twitter. They proposed a framework that uses a hybrid design to combine three kinds of classification algorithms for enhanced emoticon analysis, SentiWordNet analysis and polarity classifier. The proposed framework consists of three main modules for data acquisition, preprocessing of data and classification algorithms pipelined to classify the Tweets into positive, negative or neutral. The datasets were generated and the experiments were conducted on six different datasets consisting of random tweets acquired from Twitter using Twitter streaming API. Experimental conclusions mentioned in this chapter conclude that the hybrid methods perform significantly better than the individual components. Thus, the results show that these hybrid techniques end up with more accurate results compared with similar techniques individually applied to the datasets. They achieved an average efficiency of 85.7% with 85.3% precision and 82.2% recall while using the hybrid technique. They also significantly contributed to decreasing the number of Tweets categorised as neutral. The frameworks tested showed enhancement in precision, effectiveness and recall when hybrid techniques were used.

3 Social Network and Media

A social network is a complex system of individuals and or organisations. It is an online platform that enables people to forge social contacts and relations with other people and organisations that harbour similar interests. Today, a plethora of social-networking websites are available to internet users that provide an array of social media services such as e-mail, instant messaging, online forums, blogging etc. They often allow users to share their ideas and thoughts through comments, personalised messages, broadcast messages, public posts, digital photos, videos, and through audio. A variety of social networking websites are easily accessible to users through internet enabled desktops, laptops, tablet computers plus smartphones and

smart-watches. They keep the users up to date with real-world activities and events happening within their social network.

Since their inception, the social networking websites have amassed millions of active users. The following are some of the popular and influential social networking websites and mobile applications:

- Facebook: A social networking website and service that allows the users to stay connected with their friends and family. Users can post comments, share photographs and videos and follow pages of interest.
- Twitter: A social networking service that allows users to read and share short messages (maximum 140 characters) called Tweets.
- YouTube: A video sharing and rating website that allows user to upload, view, rate, comment on and share videos. It is the world's largest video platform.
- WhatsApp: An instant-messaging service for smartphones that enables users to send text messages, images, audio and video in real-time.
- LinkedIn: A social networking service that enables a user to build professional networks and contacts.

4 Text Mining

Text mining [3] is the discovery and exploration of hidden information by processing the textual data from various sources. The objective of text mining is to find interesting patterns from large databases. Thus, it gleans useful information from text and processes it to generate important trends through pattern learning. Text mining structures the input text by parsing and adding or removing derived linguistic features, and inserting the subsequent features into the database.

Text is the preferred method of communication in the social media. It is extremely vital to extract the information and knowledge from the profusion of text data extracted from social media and networking websites. The processing of this natural language text extracted from social media has led to the discovery of complex lexical and linguistic patterns that can assist and aid in the answering of complex questions and form an important knowledge base [8]. Thus, text mining has huge potential in the commercial and in the research and educational sectors, as the unstructured and unprocessed text is the largest easily available source of knowledge. Refined data mining techniques are needed to extract the unknown tacit and potentially useful information from the data [9].

The process of text mining can be subdivided into two distinct phases. The first phase is the text refining phase in which the transformation of free text into an intermediate form takes place. Various data mining techniques can be directly applied on the intermediate form. The intermediate form can be represented in various forms such as conceptual graphs, relational data representation and object-oriented. The second phase is the knowledge distillation form. In this phase, data mining techniques such as clustering and classification are applied to the

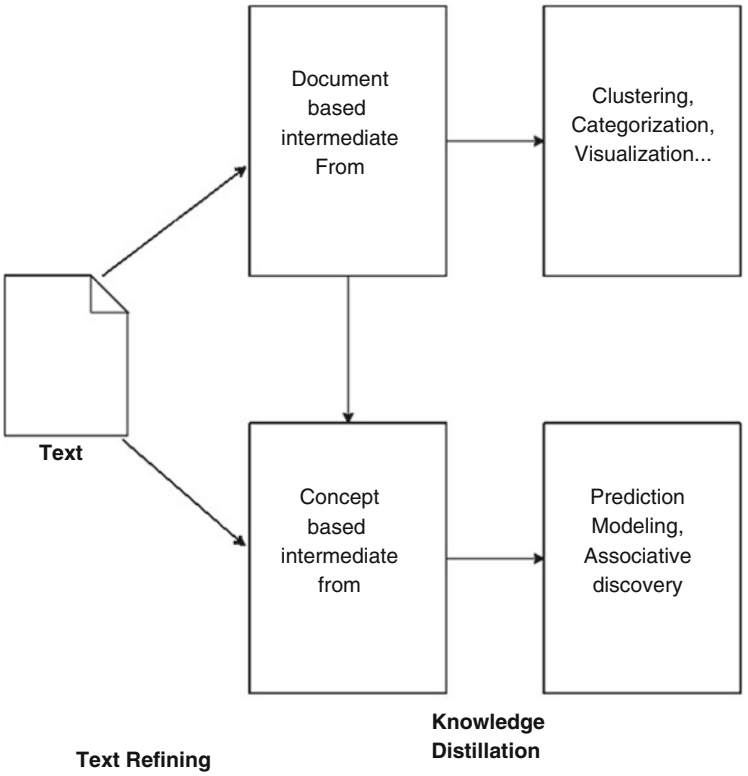

Fig. 3 General framework for text mining

intermediate form with the intention of inferring important patterns that could be incorporated into the knowledge database (refer Fig. 3) [19].

5 Opinion Mining and Sentiment Analysis

Opinion mining refers to the outlook or viewpoint of every user on certain products, services, topics, etc. Sentimental Analysis can be broadly applied to social media for the purposes of communicating, delivering and exchanging offerings that have value for the public in general and for managing the relationship between client and customer. For example, viewers post many observational comments on social media networks about news feeds, or they express their judgement and remarks on current affairs. These remarks made by each individual can be taken as opinions on the shared data that they post on these web forums [16].

The application of opinion mining would be for organisations, clients etc. to examine the reviews and evaluate the viewers and users towards their organisations

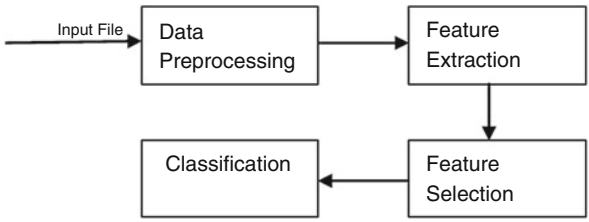

Fig. 4 Framework model for opinion mining

and the respective products. These comments regarding the company products can be given and the respective messages can be conveyed from the public to the social media sites in the form of Tweets, messages, comments etc. Several techniques of opinion mining to evaluate and scrutinise the feedback provided by the product or service users have been introduced. Several real-time scenarios, for example, reviews on a documentary for a newly released movie, form the basis of these techniques and result in increased profitability as a consequence of the research, in a commercial sense. Therefore, opinion mining forms an integral component of top companies, which are usually product-based.

The various steps of execution in the foundation model of opinion mining systems are shown in Fig. 4 [16].

6 Traditional Methods

The traditional techniques are the stand-alone techniques used for the text mining of social media data. They operate independently and cannot be sub-divided, unlike the hybrid techniques, which are made up of various composite methods. There are many traditional techniques that are used to mine the social media data. These techniques are broadly classified into machine learning-based and lexicon-based techniques.

The Machine learning (ML) techniques are mainly the supervised classification methods and therefore need two types of datasets: a training dataset and a test dataset. A supervised learning classifier uses the training dataset to learn the diverse and distinctive patterns present in the data, whereas the test dataset is used to examine the performance of the classifier. The lexicon-based techniques consist of unsupervised techniques in which the features of the texts are analysed with the lexicons having predetermined values. Some of the prominent machine learning techniques used for the mining and pattern extraction of social media data are discussed in the following sections.

6.1 Soft Computing for Web Mining

Soft computing is different from conventional (hard) computing, as, unlike hard computing, it is tolerant of inaccuracy, ambiguity, incomplete truth and approximation. It is an amalgamation of methods and techniques that operate collectively and have an impact in one form or the other on adaptable information processing for handling situations that are obscure and cryptic in real life. It is aimed at deriving benefit by utilising the endurance for inaccurate, incertitude, approximate reasoning and the partial truth to obtain tractability, resilience, relatively inexpensive solutions and close congruence to the decision-making mechanism, like that of human beings.

The principle that steers it is the formulation of the computational approach that would lead to a solution that is admissible at low cost by seeking an approximate solution to an inaccurately or accurately devised problem. There is a requirement for incorporating and implanting agent-based intelligent systems into the web tools to facilitate web intelligence. Designing of intelligent systems (both client and server-side) requires the scrutiny of the researchers from various domains such as artificial intelligence, ML, knowledge discovery etc. Such systems are capable of mining knowledge both in specific web zones and across the Internet.

Although the issue of designing automated tools to detect, extract, reduce, filter and access information that the users demand from unlabelled scattered, and assorted web data has been an unsolved mystery to date, soft computing seems like a viable option for managing these traits and properties, and for dealing with some of the constraints of existing techniques and technologies [15].

6.2 ANN for Web Usage Mining

In ML, an artificial neural network (ANN) is a nexus kindled by biological neural networks that estimates certain functions by considering a large number of inputs, usually unknown. The fundamental component of this information-processing structure is a large amount of interconnected processing components known as neurons working in unison to deal with a particular problem. An ANN is arranged in a certain configuration to work for a specific application, such as recognition of patterns or data classification, through a learning procedure.

There are various advantages of an ANN that make it well suited to web mining and analysis of social media data. An ANN is an adaptive learning technique that organises and coordinates the assignment of humans to produce the desired effect as per the requirements and demands of the users with the help of the computers. An ANN also has the ability to generate its own arrangement or depiction of the information it receives during the phase of learning. This property is termed self-organisation maps (SOMs). In addition, ANN computations can also be carried out in parallel and are fault-tolerant via Redundancy Information Coding [20].

The source of web log data taken into account for the assessment can include any specific web server for a particular time limit. Initially, data cleaning is carried out to get rid of the redundant and irrelevant log that has a potential to slow down the process of obtaining the scheme and trend of web usage. Once the data are refined, the user identification process is carried out with the help of an IP address and user's agent fields. By assisting with applying the algorithms, the users are uniquely and distinctly recognised and the path of sessions whose transactions have been accomplished are obtained.

The URL clustering is achieved with the help of K-means to obtain the frequency of each URL in each cluster. Clustering is a technique used to partition the data elements into clusters keeping similar data in the same cluster and finding unknown patterns in the datasets. Each cluster is denoted as an adaptively varying centroid (or cluster centre), beginning with some values named seed points. The distances between input data points and centroids are determined and inputs to the closest centroid are allocated K-means [4].

Steps for k-means Clustering used by Chitraa et al. [4]:

1. We achieve several transactions once the data are cleaned and the sessions are recognised. Each transition involves multiple URLs.
2. The number of input neurons to the ANN are most likely to be the pages of the website.
3. Alpha value is computed which is the threshold value, i.e. the resemblance between two transactions.
4. Choose any one transaction amongst all the recognised transactions as the centroid.
5. Select another transaction and calculate the distance between the centroid and this transaction, if the distance computed is less than alpha, it can be concluded that the second transaction is a separate cluster altogether.

6.3 Fuzzy Logic in Web Mining

Fuzzy logic can be considered a generalised form of classical logic. Lotfi Zadeh (mid-1960s) attempted to develop fuzzy logic, which was intended to assist the calculations and predictions of problems that required the use of inaccurate data or the formulation of inference rules generically using the diffusion classification.

The unit interval [0,1] is the most popular range of membership function values that are usually taken into consideration. Let μ_A represent the membership function of the fuzzy set A, which can be expressed as $\mu_A{:}U \rightarrow [0,1]$, so that for every $x \epsilon U$, $\mu_A\ (x) = \alpha$, $0 \leq \alpha \leq 1$.

Fuzzy logic consists of a multiple-valued logic, in which the values lie within the range 0 to 1, which is considered to be "fuzzy", to express the reasoning that occurs in humans. These values are termed the truth values for certain logical propositions. A particular proposition X may have 0.6 as the truth value and its complement

0.5. As per the negation operator that is applied, these two truth values do not compulsorily have to sum up to 1. Fuzzy logic can be applied as a model to interpret the NNs and their characteristics. They can be useful in determining the network specifications without the need to apply a learning algorithm [17]. Fuzzy logic is used in cases where the inaccuracy and ambiguity levels are high and to handle the variations on a more continual level. The degree of truth is the value possessed by the propositions as a mathematical model for imprecision and vagueness. Another important quantity is the membership function, i.e. a function that describes and relates the degree of membership to a value in the domain of sets in the fuzzy set.

A real-life instance may be as follows: old air conditioners used to function using an on–off system. The interpretation of each one may be denoted by a specific fuzzy set. Let an AC be at 27 °C. When the temperature rises above 25 °C, the unit is turned on, whereas if the temperature drops below 20 °C, then it might be interpreted as the AC being turned off. Fuzzy rules such as "the cooling power would be turned up a little, in case the ambient air gets warmer; the cooling power would be turned down moderately if the temperature is dropped" would be applied. The application of the aforementioned concept would simplify the functioning of the machine and would result in a more steady and consistent temperature.

The reason for using fuzzy clustering for Web mining as opposed to the traditional clustering is because the web data possess fuzzy traits and properties. The researchers have determined the way in which soft computing techniques (such as NNs, fuzzy logic, GAs etc.) can be applied to web mining as a tool to improve the efficiency of retrieval or processing of the results obtained from the search.

6.4 Ant Colony Optimisation (ACO) in Web Mining

Ant colony optimisation (ACO) is one of the algorithms used for the intention of examining social insects such as ants or bees as opposed to imitating human intelligence (Fig. 5).

The manner in which ants optimise their trail looking for food by releasing pheromones, a chemical substance produced and released into the environment affecting the behaviour or physiology of other ants on the trail is what works as a motivation behind this meta-heuristic. They, in fact, tend to traverse the path with the shortest length, which they determine with the help of the strength of the pheromones' smell on the paths visited by the other ants, which is usually the shortest, as the shortest path takes the least amount of time to traverse. Therefore, even if the same number of ants traverse both the long and the short path, the number of traversals would be higher in the short paths depositing more pheromones on the way, which would make the other ants choose the shorter path. This intelligent behaviour arising from these otherwise unintelligent creatures has been called "swarm intelligence".

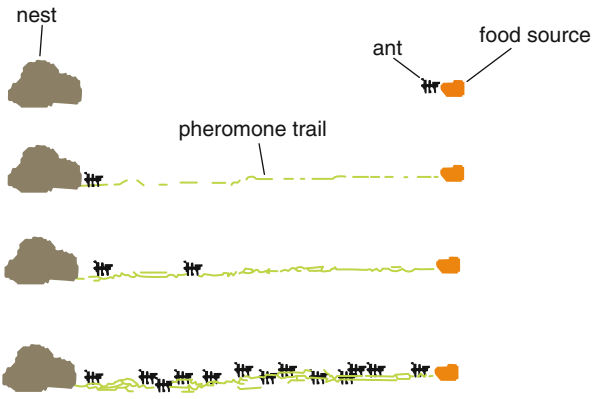

Fig. 5 Pheromone trails in natural ant colonies

Algorithm 1

Init *pheromone* τ_{ij};
repeat for all ants i: *construct_solution*(i);
for all ants i: *global_pheromone_update*(i);
for all ants edges: evaporate *pheromone*;
$(\tau_{i-j} = (1-\rho)\tau_{i-j})$

construct_solution(i) :
init *ant*;
while not yet a solution **do**
expand the solution by one edge probabilistically according to the *pheromone*;
$(\tau_{\rho i-j}/\sum_{\rho i-j^*} \cdot \tau_{\rho i-j^*})$;
end while

global_pheromone_update(i) :

for all edges in the solution **do**
increase the *pheromone* according to the quality;
$(\Delta\tau_{i-j} = 1/length\ of\ the\ path\ stored)$
end for

Fig. 6 Pseudo-code for ant colony optimisation (ACO)

This elementary concept is to the web user a chain of the web pages visited, or sessions. The ants, which have been created artificially, are made to undergo training with the help of a web session clustering technique to refashion a preference vector of the text that depicts the priorities of the users to choose the preferred group of keywords. Moreover, the behavioural pattern of browsing in the future is predicted by these ants (Fig. 6).

6.5 Particle Swarm Optimisation

Particle swarm optimisation (PSO) is a method of computation that solves a problem and obtains the most efficient solution by frequently attempting to enhance a candidate solution pertaining to the provided estimate of a quality. It considers a population of candidate solutions, the particles in this case, and provides the result with the help of the motion of the particles in the search-space by the location and velocity of the particle determined by obtaining the values using the formulas mathematically. Every particle's motion is affected by its position, which is best known locally, although it is mentored in the direction of the most preferred positions in the search-space, which are renewed as and when preferable positions are determined by other particles. This is what is intended to provide the optimal solution for the swarm.

It is derived from the idea of how a flock of birds or a school of fish behaves socially and is preferable for various applications owing to the lower number of parameters to be considered for adjustments.

7 Hybrid Techniques for Sentiment Analysis

Sometimes, two or more traditional data mining methods can be combined to yield better performance while examining social media data. Such composite techniques consisting of various traditional techniques are commonly referred to as hybrid techniques. It is often found that the hybrid techniques have an edge over their traditional counterparts as they exploit the advantages of the individual techniques to provide better accuracy and results. Some of these hybrid techniques are discussed in depth below.

7.1 Analysis of SVM-PSO Hybrid Technique [2]

The messages or certain remarks on Twitter can be used to review a movie with the help of opinion mining or sentiment analysis by applying text mining techniques, natural language processing and computational linguistics to categorise the movie into good or bad based on message opinion. In this application, the focus is on the binary categorisation, which groups the data into the positive and negative classes, with positive depicting a good message/review and negative depicting a bad one, for a particular movie that is being analysed. It uses tenfold cross-validation and a confusion matrix for the authentication process, and the precision level of the support vector machine forms the basis of the rationale. To advance the selection of the most appropriate parameter to tackle the dual optimisation problem, hybrid particle swarm optimisation (PSO) has been chosen.

The pre-existing datasets for Twitter sentiment messages are very rare. The set of data used for the research was accumulated by Hsu et al. [7]. In that, the messages

selected for the training data had emojis, which were discarded. The test data were manually obtained and contained 108 positive and 75 negative tweets from the user, which were noted manually. The dataset was then preprocessed and cleansed before being used for the research.

The result depicts the enhancement of the extent of precision from 71.87% to 77%.

One of the tiers of the flow chart (Fig. 7) is machine learning and classification, which involves the SVM. The below figure articulates the functioning of the SVM. The first step would be indexing the term of opinion ascending. Next, all the terms

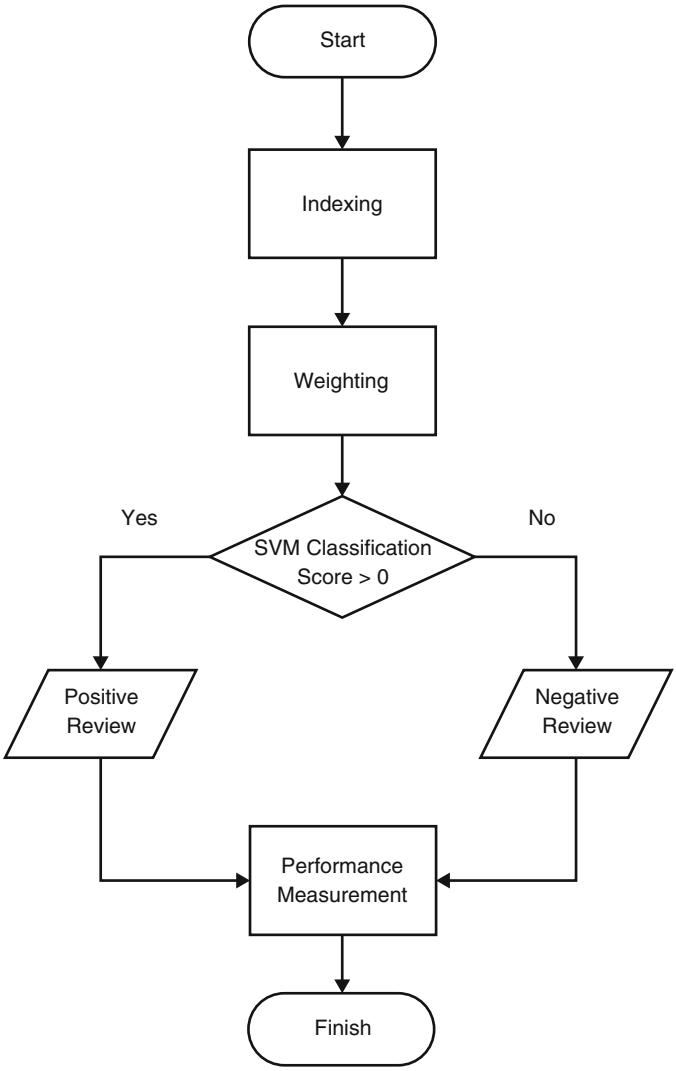

Fig. 7 Flow of the support vector machine model

are weighted according to their features. If the weighting score is more than zero (weight >0), the term is categorised as a positive review. If it is not, then the term is categorised as a negative review.

Figure 8 below illustrates the formation of PSO with population size, inaction weight, and generations without improvement. This is followed by the examination

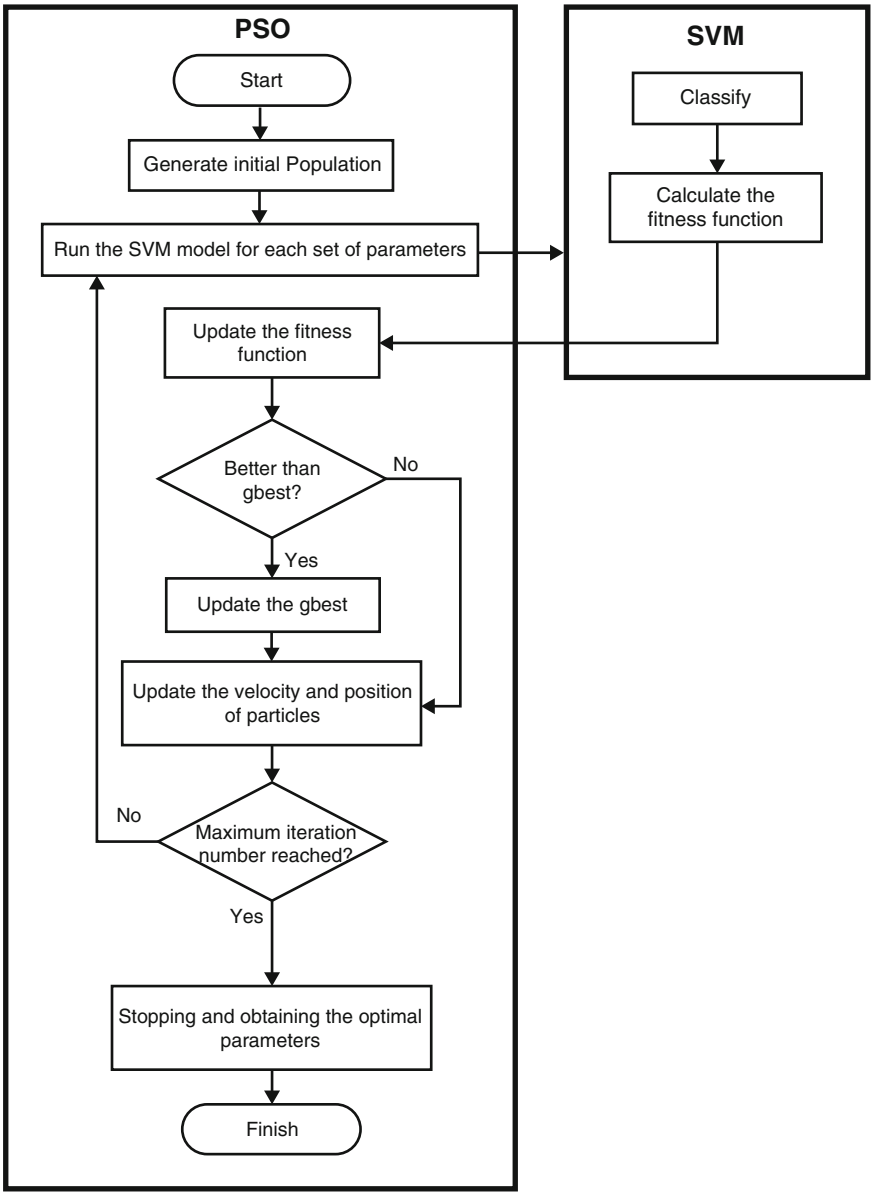

Fig. 8 Flow of support vector machine particle swarm optimisation model

of each particle's fitness. Next, is the comparison of the fitness functions and the locally best and globally best parameters. After that is completed, the velocity and position of each particle is updated until the convergence of values of the fitness task. Once convergence is finished, the SVM classifier is fed with the global best particle in the swarm for training. Lastly, the SVM classifier is trained.

The comparison was conducted among three quantities, the resultant values obtained from the SVM with the use of *N*-grams and feature weighting, the resultant values obtained by comparing the SVM with SVM-PSO without data cleansing, and those obtained with data cleansing.

The results obtained by performing the aforementioned method reflect that after the SVM and PSO are hybridised, PSO affects the performance and precision of the SVM. 77% is the optimal precision level, which is suitable for this, and that is obtained after data cleansing by the combined functioning of SVM-PSO. Although there is still potential to increase the performance with the help of enhancements of the SVM, which is mostly assisted by combining it with or varying of the SVM with some other method of optimisation (Fig. 9).

In future research, it is intended that more combinations of *N*-grams and feature weighting will be introduced that will provide an optimal level of precision compared with the above experiment. Furthermore, the categorisation of results obtained from sentiment analysis is only categorised into two categories (positive and negative). Thus, in further research, a class consisting of multiple categories for categorising the inferences obtained from further sentiment analyses (for example, positive, negative, neutral/not affected etc.) instead of just binary categorisation may be taken into account. Currently, there are no such studies in connection with social networks to show the superiority of SVM-PSO over PSO. Therefore, research can be carried out in the future to determine if the hybrid SVM-PSO technique dominates the PSO in the same way that it does with the standalone SVM technique.

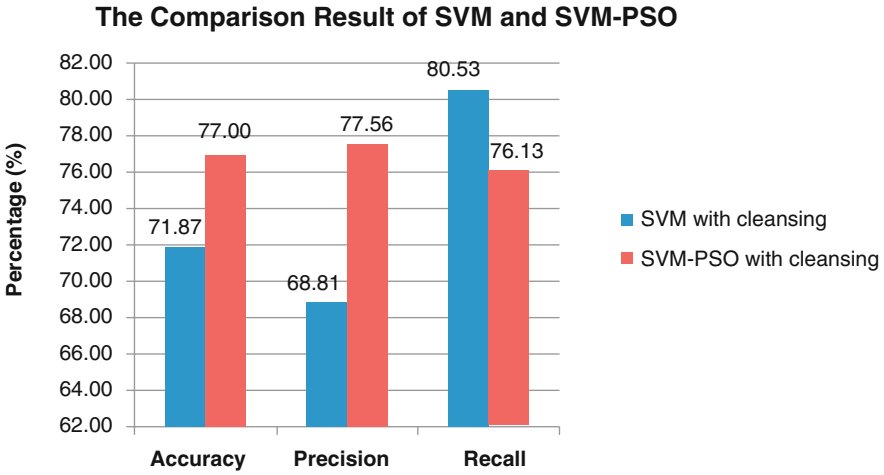

Fig. 9 The comparison result of the SVM and SVM-PSO after data cleansing

7.2 Analysis of SVM-ACO Hybrid Technique [11]

Currently, Twitter is considered one of the most prevalent micro-blogs for making short and frequent posts, which equips Twitter users to comment on a particular post, be it reviews on a movie or a book, as "tweets". These can be helpful in determining the general outlook of the public and opting for better marketing strategies. Like the previous analysis, the Tweets obtained from the Twitter data can be categorised into two sets: positive and negative, only this time with the approach of support vector machine (SVM), a ML algorithm and ant colony optimisation (ACO) hybrid strategy. The precision on an average of this categorisation increases from 75.54% to 86.74% with the application of SVM and SVM-ACO hybrid respectively.

The flow chart given in Fig. 10 depicts the procedure of opinion mining of Tweets from Twitter data. The tiers of the diagram are as follows:

1. Accumulation of data: the application programming interface (API) of Twitter offers the provision of obtaining Tweets with the keywords concerned with the help of a programmatic technique.
2. Preliminary processing: the preliminary processing of the Twitter data is done to determine and remove the redundant data elements from the input dataset. In this way, it performs data cleaning before the actual data processing to improve the precision of the categorisation. As the data obtained from the public may subsume the elements, which may not affect the polarity of the reviews, if these data are considered for calculations, it would lead to increased complications of the process of categorisation. The measures of preliminary processing involve erasing punctuation characters, such as the comma, period, apostrophe, quotation marks etc. It also incorporates a number filter to refine the numerical terms, a case converter to transform all text forms to lower case etc.
3. Generation of features: features distribute the text obtained from the source into a positive or negative category. Some of the popular methods of weighing the features include term frequency (TF), which is used for the generation of the feature vector and term frequency-inverse document frequency (TF-IDF), which are calculated with the help of Eqs. (1) and (2).

$$\text{TF} = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}} \tag{1}$$

$$\text{IDF} = \log_e \frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}} \tag{2}$$

4. Categorisation by applying SVM: A hyper-plane is generated by the SVM for categorisation (Fig. 11). The hyper-plane can maximise the nearest training instance of both categories i.e. maximise the functional margin. The prime goal is to reduce the error by generalisation and does not allow over-fitting to affect it.
5. Categorisation by applying the SVM-ACO hybrid: One of the concepts that applies swarm intelligence to tackle problems would be ACO. A trail of synchronously moving ants works coherently to deal with a sub-problem that is a

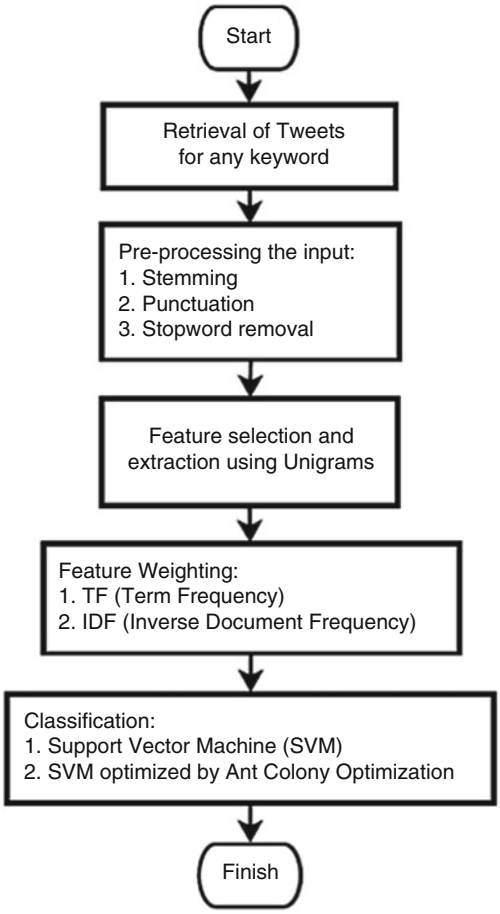

Fig. 10 Flow chart of opinion mining procedure of Tweets from Twitter data

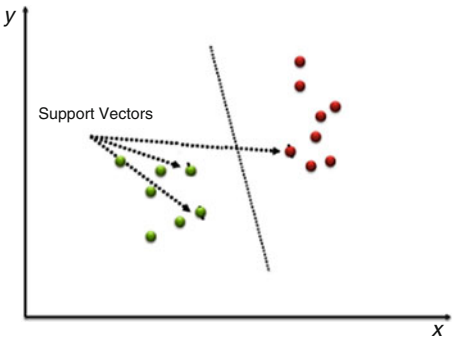

Fig. 11 Hyper-plane that segregates the two classes in the SVM: positive and negative

part of the prime problem and provides clarified solutions to this problem. Each ant builds its own solution to the problem incrementally after each movement. An ant updates its trail value according to the components used in its solution after evaluating and completing its solution or at its construction phase. The search for future ants is directed and affected to a large extent by this pheromone value (Fig. 12).

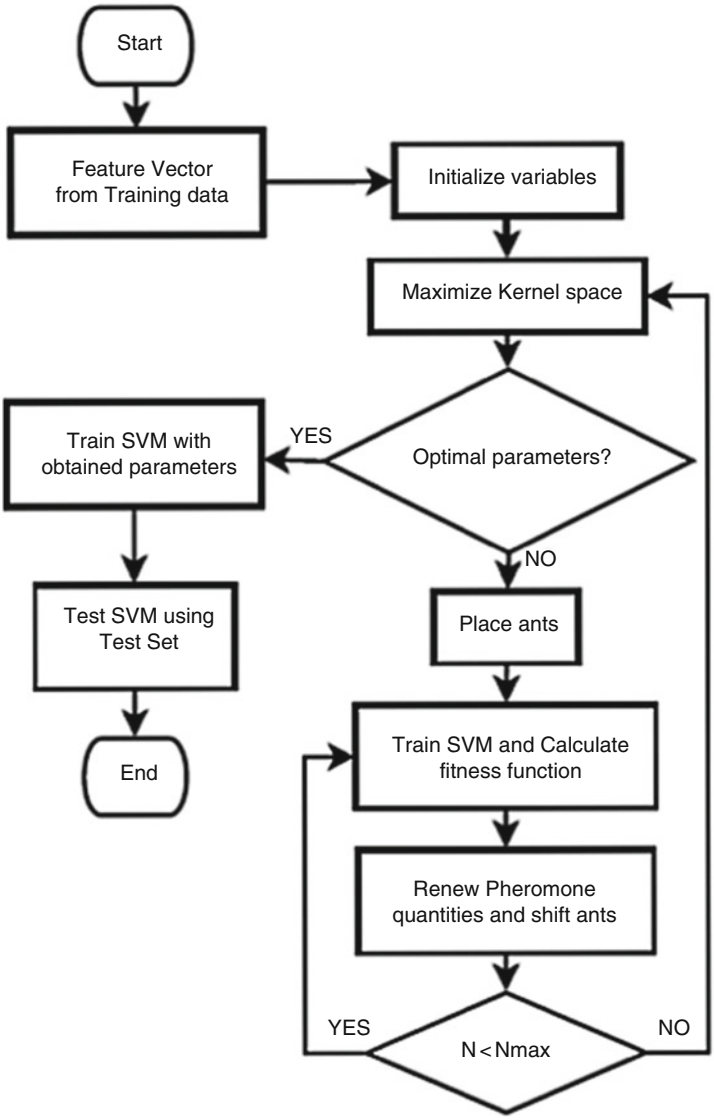

Fig. 12 Flow diagram showing how SVM is optimised with the help of ACO

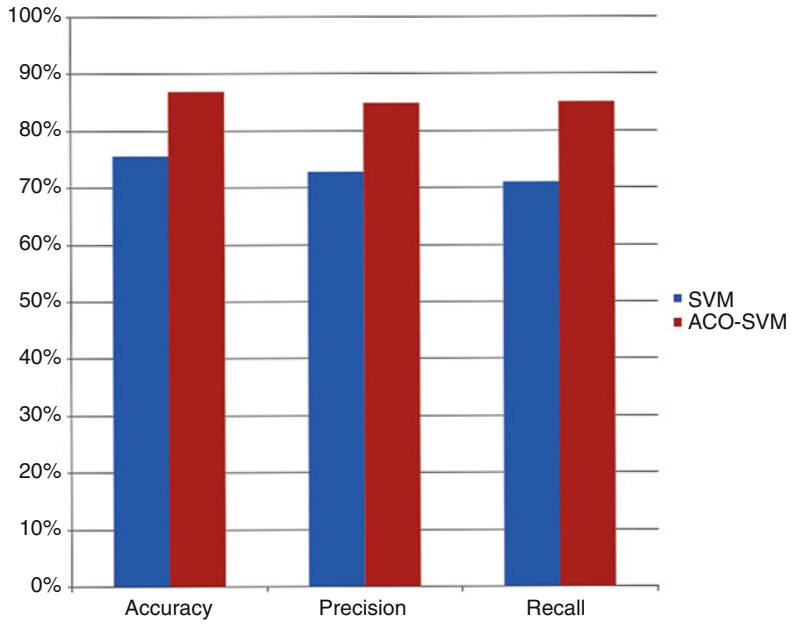

Fig. 13 Comparison of SVM and SVM-ACO

The results exhibited a huge boost in performance when the SVM-ACO hybrid technique was used instead of the SVM technique alone. The SVM-ACO technique outperformed the traditional SVM technique on all the performance parameters with an accuracy of 86.74%, precision of 84.85% and recall of 85.05%. These values were significantly better than those for the SVM, with an accuracy of 75.54%, precision of 72.77% and recall of 70.98% (Fig. 13).

There is no such study in connection with social networks to show the superiority of SVM-ACO over ACO. Therefore, there is a scope of research in this domain to determine if the hybrid SVM-ACO technique dominates the ACO in the same way it dominated the SVM technique.

8 Other Hybrid Methods

Certain hybrid methods such as neuro-fuzzy methods may enhance the speed of the web mining procedure on social media datasets. This chapter has analysed the existing techniques, methods, algorithms along with their merits, demerits and limitations, and places emphasis on potential enhancements of these methods using the soft computing framework.

8.1 Neuro-Fuzzy Hybrid Technique

Neuro-fuzzy is a term used in the field of artificial intelligence in the context of fusion of artificial neural networks and fuzzy logic. Proposed by J.S.R. Jang, the hybridisation of neural networks and fuzzy logic creates an intelligent system that coalesces and merges the reasoning of the fuzzy logic systems with the connectionism of the artificial neural network, resulting in a linguistic model that is is much more adept than both fuzzy logic (FL) and the ANN.

In Shivaprasad et al. [18], various soft computing techniques such as fuzzy logic, neural networks, genetic algorithms (GA) and ant colony optimisation (ACO) have been analysed by the authors in an attempt to evaluate their efficacy in mining the data from the social media. The research concluded that there is a requirement for fast hybrid techniques for the analysis of social media data. Hybridisation techniques such as neuro-fuzzy hybridisation can render better solutions at a reduced cost and at a faster speed. Neuro-fuzzy hybridisation is ideally suited as it has the advantages of both fuzzy logic and neural networks. It is adept in handling unclean data and in modelling the non-linear decision boundaries.

Shivaprasad et al. [18] have proposed an ingenuous hybrid neuro-fuzzy model for the mining of web data. After the collection, pre-processing and cleaning of web data, the clustering of data is carried out using the Fuzzy C Means clustering algorithm. Each cluster is unique as it comprises users exhibiting similar browsing behaviour. The artificial neural networks are constructed specific to the goals and objectives to be accomplished while imitating the architecture and information representation patterns of the human brain. In the learning process, each pattern is present in the input nodes and associated with output nodes with differential weights. The iterative process followed adjusts and re-assigns the weights between the input and the output nodes until a predetermined termination condition is met. The training of the neural network is carried out by giving the input of the clustering algorithm as the input of the neural network, whereas the output of the clustering algorithm is the target output of the neural network. The performance of the system is measured using the mean square error (MSE), which is the average of the square of error between the outputs of the neural network and the target outputs from the clustering algorithm. A multi-layered forward neural network is used which is trained with back-propagation learning algorithms.

9 Conclusion

The prevalence of social media usage has enhanced the inquisitiveness in the field of opinion mining. This has made it an essential task for several corporate institutions and companies to categorise the sentiments from the texts and public reviews in an efficacious manner. Keeping this in mind, we have tried to analyse the hybrid techniques such as neuro-fuzzy, SVM-ACO and SVM-PSO after briefly explaining

the traditional techniques and giving a brief description of all the important concepts and methodologies that are referred to in this chapter. Towards the end of each hybrid intelligent technique, the results and relevant inferences pertaining to that particular technique are taken into account.

It can be clearly observed that hybrid intelligent techniques are more efficient than their individual components. The experiments indicate that the usage of hybridised techniques significantly increases the precision level percentage. Moreover, the precision level of the hybrid methods can be further increased with the help of improvements in the existing version of the components used to generate the respective hybrid models. It could be achieved by using it in another combined arrangement or modification of the existing component models with some other method of optimisation. We have mentioned how a particular technique is suitable for a certain type of data-set and how new hybrid intelligent techniques can be created to produce better optimisation and accuracy.

10 Future Scope

- The results of various hybrid methods are examined experimentally and observationally on datasets of varying sizes for the purposes of opinion mining. Even though the hybrid intelligent techniques are more efficacious, there is a requirement and scope for development of new techniques that would be more efficient and accurate.
- In the future research on the various hybrid methods, an extended analysis is expected to be carried out to examine how further enhancement can be made in each respective area and how several domains and region-specific parameters influence the sentiment analysis results. Expanding the concept of opinion mining to various other domains may culminate in innovative results.
- A broader variety in combining the *N*-grams and assigning weight to the attributes and features that provide more precision than the existing ones can be researched and taken into account.
- The current techniques usually classify the data into two categories (binary classes) that are either favourable or unfavourable (or positive and negative). In further research, on opinion classification of more than these two categories (like an additional neutral category) is expected to be included in the results and analysis of hybrid intelligent techniques.
- There is a huge chunk of data from the health sector that can be analysed using hybrid techniques for better prediction of diseases. In future, the sentiments of the crowd can be analysed and relevant conclusions can be drawn regarding the winning party in the elections.
- Apart from social media data, the use of hybrid techniques in the medical sector also has huge potential. In their research, Mishra et al. [14] have shown that SVM-ACO can be successfully used to find the best subset descriptors for the classification of anti-hepatitis peptides.

- Furthermore, it would be expected that the various other hybrid techniques, apart from the conventional combinations, would be researched and implemented after their feasibility and analysis had been weighed up.

We consider that this chapter may also be used to obtain novel ideas for new lines of research or to continue the lines of research proposed here.

References

1. Aggarwal, C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128 (2012). doi:10.1007/978-1-4614-3223-4_4
2. Basari, A, Hussin, B., Ananta, I., Zeniarja, J.: Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Proc. Eng. **53**, 453–462 (2013). doi:10.1016/j.proeng.2013.02.059
3. Berry, M.: Automatic discovery of similar words. In: Survey of Text Mining: Clustering, Classification and Retrieval, pp. 24–43. Springer, New York (2004)
4. Chitraa, V., Selvadoss Thanamani, A.: An enhanced clustering technique for web usage mining. Int. J. Eng. Res. Technol. **1**(4), 5 (2012)
5. Feitosa, R., Labidi, S., dos Santos, A.S.: Hybrid model for information filtering in location based social networks using text mining. In: 2014 14th International Conference on Hybrid Intelligent Systems (2014). doi:10.1109/his.2014.7086206
6. He, W., Zha, S., Li, L.: Social media competitive analysis and text mining: a case study in the pizza industry. Int. J. Inf. Manage. **33**, 464–472 (2013). doi:10.1016/j.ijinfomgt.2013.01.001
7. Hsu, C.-Y., Yang, C.-H., Chen, Y.-C., Tsai, M.-C.: A PSO- SVM lips recognition method based on active basis model. In: 2010 Fourth International Conference on Genetic and Evolutionary Computing (2010). doi:10.1109/icgec.2010.188
8. Jusoh, S., Alfawareh, H.: Agent-based knowledge mining architecture. In: International Conference on Computer Engineering and Applications, pp. 526–530. IACSIT Press, Singapore (2009)
9. Jusoh, S., Alfawareh, H.: Techniques, applications and challenging issue in text mining. Int. J. Comput. Sci. Issues **9**, 431–436 (2012)
10. Kamruzzaman, S., Haider, F.: A hybrid learning algorithm for text classification. In: International Conference on Electrical & Computer Engineering, 1st edn. (2004)
11. Kaur, J., Sehra, S., Sehra, S.: Sentiment analysis of twitter data using hybrid method of support vector machine and ant colony optimization. Int. J. Comput. Sci. Inf. Secur. **14**, 222–225 (2016)
12. Khan, F., Bashir, S., Qamar, U.: TOM: Twitter opinion mining framework using hybrid classification scheme. Decis. Support Syst. **57**, 245–257 (2014). doi:10.1016/j.dss.2013.09.004
13. Kumar, R., Sharma, M.: Advanced neuro-fuzzy approach for social media mining methods using cloud. Int. J. Comput. Appl. **137**, 56–58 (2016). doi:10.5120/ijca2016908927
14. Mishra, G., Ananth, V., Shelke, K., et al.: Classification of anti hepatitis peptides using Support Vector Machine with hybrid Ant Colony Optimization. Bioinformation **12**, 12–14 (2016). doi:10.6026/97320630012012
15. Mitra, S., Pal, S., Mitra, P.: Data mining in soft computing framework: a survey. IEEE Trans. Neural Netw. **13**, 3–14 (2002). doi:10.1109/72.977258
16. Mohana, R., Umamaheshwari, K., Karthiga, R.: Sentiment classification based on latent dirichlet allocation. In: International Conference on Innovations in Computing Techniques (2015)
17. Rojas, R.: The backpropagation algorithm. In: Neural Networks, 1st edn., pp. 149–182. Springer, Berlin (1996)

18. Shivaprasad, G., Reddy, N., Acharya, U., Aithal, P.: Neuro-fuzzy based hybrid model for web usage mining. Proc. Comput. Sci. **54**, 327–334 (2015). doi:10.1016/j.procs.2015.06.038
19. Sumathy, K.L., Chidambaram, M.: Text mining: concepts, applications, tools and issues an overview. Int. J. Comput. Appl. **80**, 29–32 (2013). doi:10.5120/13851-1685
20. Verma, P., Keswani, N.: Web usage mining: identification of trends followed by the user through neural network. Int. J. Inf. Comput. Technol. **3**, 617–624 (2013)
21. Zafarani, R., Abbasi, M., Liu, H.: Social Media Mining: An Introduction, 1st edn. Cambridge University Press, Cambridge (2014)
22. Zhai, C., Massung, S.: Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, 1st edn. ACM Books, New York (2016)