Blog Response Volume Prediction using ANFIS and Stochastic Optimization Techniques

Thesis submitted in partial fulfillment of the requirements for the award of degree of

Master of Engineering

in

Computer Science and Engineering

Submitted by

Harsurinder Kaur

(Roll no: 801632009)

Under the supervision of

Dr. Husanbir Singh Pannu

Assistant Professor, CSED

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY
PATIALA 147004

June 2018

Certificate

I hereby certify that the work, which is being presented in the thesis, entitled "*Blog Response Volume Prediction using ANFIS and Stochastic Optimization Techniques*", in partial fulfillment of the requirements for the award of the degree of Master of Engineering in *Computer Science and Engineering* and submitted in Computer Science Department of Thapar Institute of Engineering and Technology (Deemed to be University), Patiala is an authentic record of my own work carried out under the supervision of *Dr. Husanbir Singh Pannu* and refers other researcher's work which are duly listed in the reference section.

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

Harsurinder Kaur
Signature

Harsurinder Kaur

This is to certify that the above statement made by the candidate is correct and true to the best of our knowledge.

Hs Pannu

Dr. Husanbir Singh Pannu
Assistant Professor,
CSED

Abstract

Due to wide streaming blogs over the social media, blog volume automation has become indispensable for the analysis of blog popularity. As a rule base driven method, Adaptive Neuro Fuzzy Inference System has gained popularity in various prediction tasks for its efficiency and ease of implementation. In this paper, two modified Adaptive Neuro Fuzzy Inference models have been proposed by tuning its premise and consequent parameters with Particle Swarm optimization and Genetic Algorithms. Particle Swarm optimization helps in reducing the training and cross validation error of the predictive model whereas Genetic Algorithm optimize minimum clustering radius which aids in the formation of rule base. With the help of these optimization methods, Adaptive Neuro Fuzzy Inference System obtains optimal premise and consequent parameters which improves its predictive performance. Comparative analysis of proposed method has been performed against Neural Networks, Support Vector Machines and basic Adaptive Neuro Fuzzy Inference System using UCI Blog Feedback dataset. It has been found that both of the proposed variants have outperformed these state-of-art techniques.

Acknowledgement

First, I would like to extend my deep gratitude to my supervisor **Dr. Husanbir Singh Pannu** for constant supervision, for their advice and patiently guidance at every step of my ME program. Without unfailing support and perception in me, this thesis would now not have been viable. Their contribution to this thesis goes well beyond their role as an academic supervisor and includes constant support on a personal level without which this journey may never have been completed. And for this, I'm truly thankful. He is a great mentor for my life as well.

I would like to express my gratitude to **Dr. Maninder Singh**, Head of Computer Science and Engineering Department and **Dr. Ashutosh Mishra**, P.G. coordinator their constant motivation and encouragement.

I also wish to thank my research committee members and non-teaching staff of the Computer Science and Engineering Department for their help and support. I would also like to thanks to my teachers and friends from whom I learn the art of happiness and never give up approach.

Finally, I would like to express my sincere and deep gratitude to my parents and family members for their love, encouragement, care, and support.

Harsurinder Kaur

Table of Contents

List of Figures

List of Tables

List of Notations

$O_{l,n}$	output of nth node at layer l
$O_{i,n}$	membership value of fuzzy set
a_i, b_i, c_i	premise parameters
p_t, q_t, r_t	consequent parameters
$X_i(t)$	ith particle position at time interval t
$V_i(t)$	ith particle velocity at time interval t
G_best	global best
P_best	personal best
$X_i(t+1)$	ith particle position at time interval t+1
$V_i(t+1)$	ith particle velocity at time interval t+1
$\mu_A(x)$	Guassian function
A_1, A_2, A_3, A_4	fuzzy set

List of Abbreviations

ACO	Ant Colony Optimization
ANFIS	Adaptive Neuro Fuzzy Inference System
AUC	Area Under Curve
BBC	British Broadcasting Corporation
CNN	Cables News Network
DT	Decision Trees
EA	Evolutionary Algorithm
GA	Genetic Algorithm
GDA	Generalized Discriminant Analysis
HDP	Hiearchical Drichlet Process
ICA	Independent Component Analysis
k-NN	k-Nearest Neighbor
LDA	Linear Discriminant Analysis
LR	Logistic Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MART	Multiple Additive Regression Tree
MSE	Mean Squared Error
MLP	Multi Layer Perceptron
NB	Naives Bayes
NN	Neural Networks
PCA	Principal Component Analysis
PSO	Particle Swarm Optimization
RBF	Radial Basis Function
REP	Reduced Error Pruning
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
SVR	Support Vector Regression
TF-IDF	Term Frequency-Inverse Document Frequency

UCI	University of California, Irvine
UCM	User Centric Model
URL	Uniform Resource Locator
URM	User-Retweet Model

Chapter 1

Introduction

The work presented in this thesis covers the prediction of volume of response in blogs. So, for the prediction of the response volume, a new technique using Adaptive Neuro Fuzzy Inference System combining stochastic optimization methods is proposed. This chapter introduces with various terms associated with blogs and application areas of prediction subjects in blogs.

With the ever growing advancement in use of technology in the era of digital world, social media has shown incredible importance in lives of people since last decade. It has now become a fact that assuming a life without web seems impossible. While in early period of development of social media, blogs, tweets, YouTube was just an source of entertainment for few passionate users. But now it has become a goldmine for building the relationship between the users and following their interest. As, it has lead to change in the way we communicate, share our experience and opinions, and proves to be a rich source for education [1], marketing [2] and government organizations [3]. With getting knowledge about world, online socializing on web, people have discovered ideas to effectively use web in their daily lives. Also, it has become easy for the companies for promoting their new products, services through advertisements and posts over social networks. While positive responses over blog posts build their popularity and hold over social media, on other hand, negative comments and reviews often affects the ratings of companies. As negative responses catch attention of people much rapidly than positive ones and companies might need to take quick actions for avoiding the fall.

Basically, blogs refers to website which is created by user in which text is added for readers which is reflected in a reverse chronological order on the basis of created date. Blogs are usually created by single person or group of persons who are having the same interest and provides two way communication. Blog entries usually reflect the blogger's interest or views about a topic streaming around and it can be in the form of discussion over some topic with other bloggers. Blog websites supports

Figure 1.1: Users comments

views points in the form of comments from the readers links to other related blogs and reference to own older posts which is known as track backs. For example, on bloggers websites such as WordPress [4], Blogger [5], readers can input comments which are descriptive opinions on the blog posts. General example of comments by the users is depicted in Figure (1.1). Comments in the blog posts generate quality traffic.

The shift in the popularity of blogs usually depends on the news streaming in society or discussion over a hot topic. These type of posts attract new readers to acquire knowledge. Sometimes, these posts turns into debate or aggressive views which fades with the arrival of new topic. This maintains the life cycle or longevity of a blog. Eventually, this turns into most popular blog looses their popularity often with time which becomes a subject of research for predicting the popularity.

1.1 Terminology of Blogs

Besides blog forms are ever changing for making it reliable for the readers and analyzing their interest. The description of blogs and various terms into blogging sites are explained as follows:

Blog post: It refers to the record of the documents published known as posts. Every blog generally has a heading which represents the concise description about the subject of the post. It also includes the date of publishing of the post, subject and author's name. In general, newly added posts always gets space at the top of

the website which represented in a reverse chronological order according to date [6].

Blog space: It is also referred as blogo-sphere, is confined with all the blog posts and their interconnections, which supports concept of blog as a social world [7]. Blogs are often become popular when it contains various hyperlinks making a connection between so many blog posts. It helps in improving the life cycle of the blog as associated links often improves popularity.

Blogger: it refers to user who is creating blogs for the people. Blog posts are often derived by the interest of the blogger and their inspirations. Blogger's network improves the profile. Blog publishing on websites makes difference in the popularity of the blog posts.

Trackbacks: It refers to the idea of linking pages to the related posts, in this context it refers to link connection between related blog posts [8]. While clicking the links, readers can access the content of the referenced blog posts. It also helps in improving the lives of multiple posts.

Comments: It refers to view points or reactions of the readers over the content of blog which is represented after publishing it. Most of the blogging websites such as Blogspot or Blog Tyrant accepts comments over the blog posts. Number of comments enhances the popularity of the blog post [9]. It appears in the order that last comment is usually to be appeared at last.

Permanent Links: It refers to as address of URL connecting to a specific blog post [10]. The blogs websites referred by these URLs are constant so that it can be accessed every time user clicks on it.

Blog rolls: Blog rolls is known as the entries added by the blogger on its own profile. These are the most effective blog posts which describes the interest and finding of the blogger. These are also the posts which are mostly liked by the people.

1.2 Prediction of feedback in Blogs

In blog document most relevant contents which need to be considered are main text, feedback and links to other documents. The popularity of a blog post based on these three aspects is illustrated in Figure (1.2). For bloggers maintaining a

wide network of individuals is a challenging task. Over an average of 50 billion posts are being published on daily basis which is ever increasing [11]. The evolution of multi-author blog which are published by group and edited by experts has increased the traffic of blogs posts over web. So, due to wide streaming posts and blogs over social media daily, analysis of attitude and thoughts of users is considered to be a important and challenging task. Feedback of Blogs means getting responses from the people after publishing it over social media, it contains the view-points and thoughts of people for the subject of the blog. It brings the people of same interest together and they interconnect with each other. Feedback of blogs is collected in the form of likes, dislike, comment and share. Feedback of people is dependent on the subject, content, time, type, interest, number of followers associated with that, profile of the blogger. Even feedback from blogs is connected with profile of the bloggers, their posts and network [12]. Prediction of blogs is important because with the analysis of behaviour of people, it is easy to check over the interest and subjects attracting the followers. This idea can help in enhancing the popularity of the blogs. Even, new bloggers can analyze the patterns of users and can effectively choose every key idea to become famous in very short period of time. With this we can check the attitude of users over particular issue. This also helps bloggers in sentiment analysis [13] of users. The most widely used prediction subjects in which blogs are mostly published along with the examples is described in Table (1.1).

Due to immense volume of posts sounding over social media, it has become difficult for the analysis of these post by the human experts. But there are so many social networking sites which has numerous users and wide changing patterns of the users and their behavior, asks for the automate analysis of posts over social media on daily basis. Moreover, information over social media is uncontrolled, ever-changing which arises difficulty for the manual analysis of posts and information. For example, when a post appear on social media, users may rapidly respond to that post.

In the analysis of responses of posts over web, opinion mining has gained a lot of attention. Thus, opinion mining has been taken into consideration while comparing the proposed prediction process with the opinion mining which is explained in

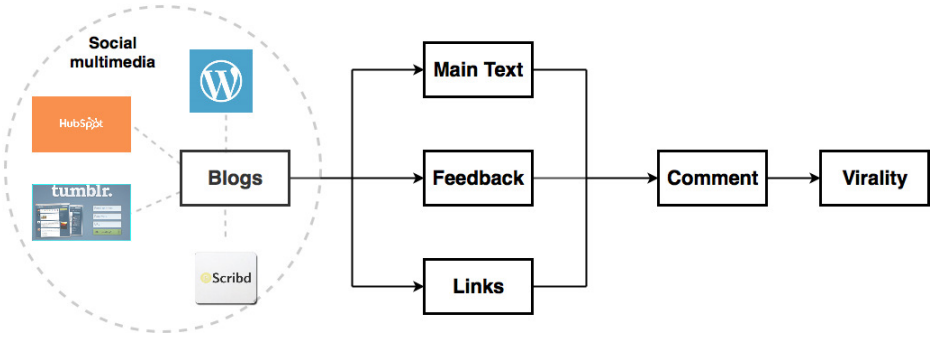


Figure 1.2: Illustration of Blog response

the next section.

1.3 Opinion mining in blogs

Opinion mining is performing a great role in this modern era. Basically, opinion mining is field of research where automated analysis of human opinion extracted from the textual documents scripted in natural language. Analysis of opinion is important because it leads to better understanding of interest of users, their attitude while reading the published documents. For example, view points of users are extracted from movie reviews in terms of quality. It has application in text mining, natural language processing, statistics and machine learning. Although there are so many tools for automatic analysis of opinion from text such as web fountain [14], Red opal [15] and opinion observer [16]. Basic elements of opinion mining are as follows:-

1. *Opinion*: It is considered as view point, attitude or suggestion for the object created by user [17]. It can vary from person to person and reflects the interest of people in specific object.

2. *Opinion holder*: It is the reviewer who provides specific response for a post [18].

3. *Object*: Considers the entity for which view points are shown.

4. *Feature*: Refers to components of object based on that opinions are expressed

5. *Opinion polarity*: It refers to evaluating the sentence whether it expresses positive, negative or neutral opinion towards the sentence.

For predictive analysis of view point(s) machine learning is used.

Opinion mining is applicable at various levels in extracting the text which are described as follows:-

1.3.1 Document level

It refers to subject oriented level in which opinions are expressed such as positive, negative or between positive or negative i.e. neutral [19]. It classifies documents on the basis of overall opinions gained by the readers or opinion holders [20]. The details of the interest of the opinion holder is not present in this level.

1.3.2 Sentence level

This level refers to each sentence of the document and evaluates the polarity of each sentence [21]. It is related to opinion mining at document level. It analyze the the sentence by two objectives which are by first evaluating the subject related or object related opinions from that and then analyzing the opinions into positive, negative and neutral and produces a summary report [22]. It also accounts for collaborating feature synonyms.

1.3.3 Entity & Aspect level

It refers to the main and important level of the opinion mining. It considers the features of specific object, in which features are component and attributes representing the object [23]. It evaluates the features of the object on which comments have been added and then reports the opinion as positive, negative and neutral. It has disadvantage that it is not relevant for opinion mining in blogs.

But opinion mining or sentiment analysis techniques often analyze feedback into feature vectors on the basis on word, document and sentence level and then classify into positive, negative and neutral opinions, which has some limitations such as huge amount of data is required, more Time(s) for analysis, language barrier. It works for finding the opinions of the people. Thus, it cannot be done only on the basis of single data point, it needs aggregate of data. Often mining from large and making data relevant is time consuming process. Even the sentiment of the

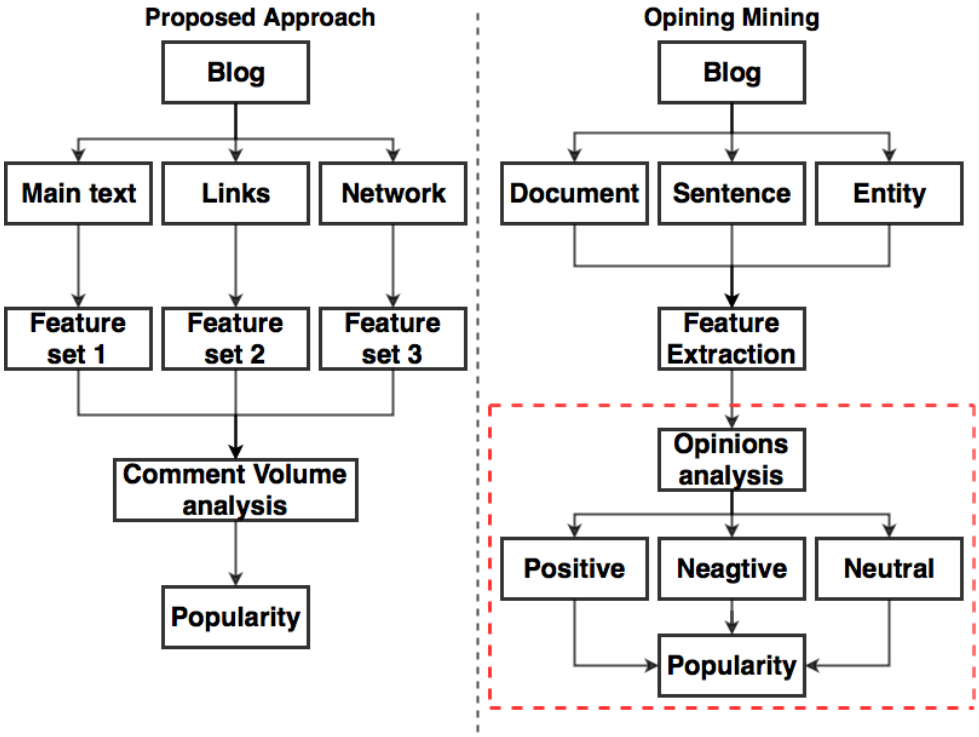


Figure 1.3: General idea of proposed approach versus opinion mining

people usually changes over time and mood, which shows a great difference of performance of the sentiment analysis. Spam and inauthentic conversations are hard to examine out of the wide data streaming and increasing on daily basis. While there are websites which only count the number of responses rather than opinions classification. Thus, in contrast general concept of the proposed approach and opinion mining is depicted in Figure (1.3).

So, in this work, automatic analysis of responses in the form of comment of blogs has been done using machine learning approaches to count the number of responses a post will receive in next 72 hrs after publishing and analyzing the trends of first 24 hrs after publishing. This approach has been performed because some users just focus on the number of hits and clicks they gain, as a metric to gauge their popularity. They just monitor the clicks over regular periods of intervals to statistically observe the increase or decrease in the user attractions. Thus, opinion mining could be unnecessary and overwhelming venture for them since it may be immaterial what sort of popularity the underlying website is gaining [24]. So, an alternate workaround has been proposed in this study, to forecast the number of

comments to quantify the website popularity. Our research is directed towards predict the volume of comments a post will receive after publishing in next H hours using modern machine learning regression algorithm.

Table 1.1: Application areas of blog response prediction

Prediction Subjects	**Application Areas**	**Related Example**	**Related Articles**
Business & Management	forecasting reviews of products, brand awareness, strategies, behaviour of people towards new products, communication with customers, topic specific comments	Apple iPod music player & Starbucks	[25], [26], [27], [28], [2], [29]
Elections	political orientation, people response towards electing new president during votes, awareness of agendas, strategies	Electing President Donald Trumph & Narendra Modi	[30], [31], [32], [33]
Public Health	online health information, individual doctors blogs, awareness about diseases, medications reviews	blog article on breast cancer, depression, allergies	[34], [35], [36]
Government & Public	web-based interactions of government with citizens, collaborative decision making, predicting responses about agendas, laws enforcement, public opinion	e-government models, US Federal Government, elected officials blogs	[37], [38], [39], [40], [41]
Social Organizations	users behaviour towards social issues, fund raising appeals, visual stories, attracting volunteers, prediction for fund raisers, users engagement in events	UNICEF, Khalsa Aid official blogs, blogs of non-profit social workers	[42], [43]
Trending News	number of comments on a news article, people varying reactions to news, blog news, awareness about hot trending topics	News blog articles on Twitter	[44], [45], [46], [47]
Icons	number of followers and haters, gossips and popularity of famous personalities, presenting their daily routines	Facebook Pages, official blogs on WordPress	[48], [7], [49]

Recipes	responses on new recipes, popularity of indigenous recipes in other regions, classification of recipes, advertise cooking classes, restaurant reviews	Official blogs of bakers, chefs and home makers	[50], [51]
Tourism	interest analysis, opinions about recently visited, ideas for new places, traveling blogs with reviews	Official blog websites such as Bucket List Journey, View From the Wing	[52], [53], [54]
Education	online learning from blogs, quality of content, analysis of interest, research based blogs, popularity of topics, new methods of learning, educational issues	E-learning blogs such as Digital Chakie	[55], [56], [57]

1.4 Thesis Organization

This thesis is organized as follows:

- **Chapter 1**: This chapter introduces the blogs, feedback in blogs, terminologies in blogs, opinion mining in blogs, levels of opinion mining, difficulties in analysis of blogs, application areas of blog prediction. This chapter also considered the conspicuous research gaps in the area of analysis of blogs. Further, research objectives are specified within the research methodology included in this study.
- **Chapter 2**: This chapter includes the latest research done in the prediction of volume of responses in blogs using various techniques. The recently used techniques for prediction of responses includes Support Vector Machine, Neural Networks, Decision Trees, Naives Bayes and Logistic Regression.
- **Chapter 3**: This chapter covers research gaps, problem statement and contribution. It defines the gap in the area, its framework and solution proposed in this study. The contribution of this study is described in points.

- **Chapter 4**: This chapter defines the proposed technique of combining ANFIS with stochastic optimization methods such as PSO and GA. Block diagram of tuning parameters of ANFIS with PSO is shown and discussed. Another block diagram showing ANFIS parameters tuning with GA is also defined. Geneal idea of ANFIS and its five layers, PSO and GA are also included in this section.

- **Chapter 5**: This chapter focuses on describing dataset and the important feature set derived from the data are defined and shown. Data sample is also depicted. It also defines PCA and its graph showing the information gain. Evaluation metrics are also discussed in this section which includes accuracy, correlation, mean squared error and time. Results drawn are described in tables. Figures containing graphs are also drawn showing the results obtained using ist principal component and while using two principal components under the light of all evaluation measures.

- **Chapter 6**: This chapter concludes the research done in this thesis and limitations found are also described. Future extension to this area is also described in this.

Chapter 2

Literature Survey

This chapter analyses the work done by various authors in the area of prediction of response volume in blog using various approaches.

In [58], Krisztian Buza predicted the number of feedback a blog article receives after publication. The aim is to model samples recently added blog articles and predict the number of response a blog article will receive in next H hours. The articles have been extracted to find the four features, basic, textual, weekday and parent features and used seven predicted models which are neural networks, RBF networks, regression trees, nearest neighbors, multivariate linear regression, bagging out of ensemble and SVM on Hungarian blog sites data. Comparison of results are carried out on two evaluation measures used Hits@10 and AUC@10 which represents the 10 blog pages which were predicted to be largest number of feedback and which have largest number of feedback in reality, respectively.

In [48] authors have designed method and apparatus for forecasting community responses to a post over social media. The number of responses are measured by using prediction model which analyze number of reactions, number of users, life of a post and generates sentiment value.

Bin Be at al. proposed two Bayesian non parametric models, URM (User-Retweet Model) and UCM (User Centric Model) which analyze user behaviour on Twitter, based on tweet text and re-tweet structure [59]. These models aids to infer the subject of choice to an respective user and predict unseen new data. Twitter real data was used to compared both models to other method, HDP(Hierarchical Drichlet Process). The described models outperforms HDP in terms of concluding topics according to interest and predictive powers. Among both models, UCM has great ability to appropriate modeling of re-tweet structure and outperforms URM in terms of overall perplexity.

Forecasting negative links over social network data is proposed in [60]. Forecasting is done by keeping positive links and subject centered interactions using NeLP

framework because these sources are easily available on social media. Prediction is analyzed on two real world social media datasets, Epinions and Slashdot by using precision and F1-measure as evaluation metrics. Experimental results shows that proposed technique can correctly predict negative link and can aid other applications also.

In [61], authors proposed the detection and classification of user behaviour based on the comments over Facebook by integrating real-time sentiment text analysis and batch data analysis for opinion mining. For evaluation "United States Presidential Elections" 2016 subject was used. 200 posts were collected from two famous channels, BBC and CNN Facebook channel. Results depicts that how people's behaviour changes over time towards a topic and represents three types of people's opinion over a real time stream. Results are evaluated on the base of Mean Absolute Error(MAE) for prediction.

In [46], prediction of longevity of online news articles based on the reaching patterns and responses of the users is proposed. It shows that detecting first 10 to 20 minutes of responses can support accurately predicting the future responses news article will gain. Linear regression models have been used to predict the number of responses a news article gain after created.

[47] proposed forecasting of popularity of news article by analyzing the number of comments by the user. The prediction is done by linear prediction model using real-world dataset gained from 20 minutes website and comments collected over the period of 4 years. The evaluation metrics used are accuracy for observing the popularity of article during short observation time and correlation coefficient for matching the actual and calculated results. [41] explained the dealing between the content of political blog and prediction of number of response or comments it will receive. Authors have motive to find out the blog posts which derives a massive response to analyze the interest of the readers. They have used two blog datasets, Yglesias and Redstate blogs on which they applied five and three predictive models respectively. Topic-Poisson model outperforms Naives Bayes in terms of recall while predicting on large volume posts and designed patterns of interest of users.

[7] explained the importance of comments in the blogs and popularity associated with the number of comments for a blog post. They have examined approximately 685976 web bogs on the basis of disputive and non-disputive comments from the web blogs. Experimental results showed that comments extraction from web blogs is evaluated on the basis of two classification using precision, recall and F-score which showed overall accuracy of 0.88.

[49] proposed classification of temperature of blog articles into four categories explosive, hot, warm and cold which are indicators of the popularity of the blog articles in the decreasing order. They have examined two data with political discussion blogs from Korea, which are SEOPRISE and AGORA. Predictions are calculated on observing the trends of response in half an hour after published using hit count and saturation point as evaluation measures for predicting the temperature.

[62] proposed the wide analysis of user comments on YouTube by examining massive amount of comment data gained by using YouTube data API. Data Mining methods have been applied to analyze the comments after dividing discussion posts, inferior comments and substantial comments. They have proposed that analyzing the trends of comments by user can help in deriving better search results.

Table 2.1: Summary of recent techniques

Paper	**Methods used**	**Idea**	**Data**	**Outcomes**	**Limitations**
Florian et al. 2017 [63]	NN	prediction of emotions of Facebook reactions	posts of supermarkets	MSE (0.135)	domain specific application
Udeen. (2015) [64]	NN and Adaboost	classification of comment	UCI blog feedback data	Adaboost outperformed with accuracy (91.4%)	features are not extracted
Singh et al. (2015) [65]	ANN and DT	Facebook pages response volume prediction	Hadoop cluster	DT-AUC@10(6.7) & RBF network-Hits@10(0.945)	time consuming regression algorithms

Paper	Methods used	Idea	Data	Outcomes	Limitations
Hieu et al. (2016) [61]	clustering techniques	user interest and attitude analysis	BBC & CNN Facebook posts	MAE (0.008)	training time of model
Chavan et al. (2015) [66]	TF-IDF, N gram, SVM, Logistic regression	response classification into bullying or non-bullying	Kaggle Public dataset	AUC-(86.9%)	only aggressive and non aggressive comments are used, comments like sarcastic or double meaning are missing
Allessandro et al. (2016) [67]	multinomal logistic model	comments classification on posts of Facebook & Youtube	Fb Graph API & Youtube Data API	Accuracy 0.80	video specific domain
Jinghua et al. (2015) [68]	CNN	prediction of popularity of dress images	Pinterest & Polyvore websites	Pinterest -Accuracy (0.84)& Polyvore-(0.83)	less feature sets are examined
Roja et al. (2012) [45]	LR, SVR, DT, SVM, NB	comments forescating of news article, classification of article publication on Twitter based on fetaure selection	44,000 news articles	R-squared value, LR (0.34), SVM (0.32), SVM (81.54), NB 77.79, bagging 83.96, DT 83.75	more overlap between features because of interdependencies
Elaheh et al. (2013) [69]	LR, SVM, DT, NB	Feature extraction for filtering useful comments, classification	Flickr comments data	Accuracy 89%	feature set based on polarization is missing

Paper	Methods used	Idea	Data	Outcomes	Limitations
Elaheh et al. (2013) [70]	LR, NV	characteristics of comments, feature extraction and classification into useful or non useful comment	Flickr, YouTube	Flickr Precision 0.87, Recall 0.90, Youtube 0.65, 0.83	user related features are not considered
Yoav et al. (2012) [71]	MART, maximum entropy	response prediction of micro blog	Twitter Firehose	MART outgrows maximum entropy classifier	unable to extract language specific features, social network of users
Dongwen et. al (2015) [72]	SVM^{perf}	cluster extraction using word2vec, classification of comments	comments over clothing product	accuracy 90% on lexical features	SVM^{perf} inability to handle multi dimensional data, only two feature selection methods used
Krisztian et al. (2014) [58]	NN, SVM, RBF net, LR	comment volume prediction	Hungarian sites document	basic features are the most predictive and siginificant	SVM inability to cope with large data
Yan Ying et al. (2014) [73]	correlation	prediction of comments over visual data	SentiBank data	yielded more accuracy then only using publisher affective content	domain specific technique
Chiao-Fang et al. (2009) [74]	SVR	ranking of comment on basis of importance and preference	Digg stories	spam and un-useful comments detection	more characteristics of comments are required

In [75] authors have forecasted box office using microblogs. Firstly, feature extrac-

tion has been done on the the micorblog data, which is from Tencent Microblog and Box office dataset is extracted from Entgroup box office website. Comments have been divided into three categories which are beneficial, harmful and neutral. Machine learning models SVM and neural networks have been applied on the data and evaluated using correlation and average MAPE.

In [65] authors have proposed forecasting volume of responses of a leading social networking service, Facebook. They have used data for prediction from Hadoop clusters and forecasted using regression models, Neural Networks (NN) and Decision Trees (DT) in contrast of other models such as Multi Layer Perceptron, RBF networks, M5P and REP trees. They have performed four data pre-processing modules and divided it into four feature sets such as Page, essential, weekday and basic features. Evaluation parameters used in this work are Hits@10, AUC@10, Mean Absolute Error(MAE) and time. They have concluded that decision tree outperforms other regression models in all evaluation measures.

Prediction of news articles gained form CNN news data is performed for analyzing the article on the basis of user opinion in the means of comments is proposed in [76]. Features are extracted on the basis of POS and frequency. In this study, Random Forests (RF) outperformed Decision Trees (DT) and K-Nearest Neighbor (K-NN).

In [64] discusses binary classification of comments on a blog post. Classification is done on online available benchmark feedback blog dataset using Neural networks and Adaboost algorithms. Extracting the four feature sets from dataset, based on that prediction rate of both the algorithms has been considered. Moreover, computational cost of both algorithms had been evaluated on all feature sets. Experiment results shows that Adaboost classifier outperforms neural networks in light of all the evaluation measures with predictive rate of 91.4%.

In [74], authors have proposed ranking of comments on the basis of its significance using machine learning algorithm SVR. Ranking of comments helps in detection of spam and un-useful comments, filters the comments with high order and preference. Feature of comments has been selected on the basis of User network, readability and content using comments on Digg Stories.

Prediction and determination of thoughtful comments is examined in [77] using logistic regression algorithm. At first, features have been selected into 7 feature sets on the basis of syntax, lexical features and discourse. For this evaluation, data from two political speech has been generated which are Singapore PM's speech in 2010, other is US President speech in 2011. After extracting the feature sets on the same basis in two data, correlation coefficient is measured on in the US and Singapore data which resulted positive. F-measure, recall and precision evaluation measures have beenn used which shows that with bagging of feature sets F-measure score gained 78.69.

Viewer affective comments form social visual data in the form of images is forecasted in the paper [73], known as visual sentiment analysis. For analyzing the publisher content comments 400 images have been used from social media. Prediction is examined on the basis of finding relationship between the the publisher content comments and then comments on the images. Thus, statistical correlation is calculated between the viewer affect concepts and publisher affect content so that relationship can be analyzed. For calculating correlation, public SentiBank data has been used which is extracted into emotion keywords and dividing into positive and negative comments. It yielded 20% more accuracy in visual sentiment analysis as compared to when only predicting comments using publisher affect contents.

Prediction of offensive attitude from tweets against the topic of USA Presidential election 2016 is performed using semantic analysis and then classification of the data using supervised machine learning techniques including NB, SVM and logistic regression. In this, SVM outperformed other methods with 87.43% accuracy [78].

Forecasting of comments generated by people on social media websites is performed in [70]. Authors have also examined characteristics of the comments, factor dependencies of comments and variations in the comments. Feature are extracted on the basis of author and social level, semantic and topic level and text and linguistics features and thus classified into useful and non-useful comments. Social data has been used for prediction from social websites such as Flickr (comments on

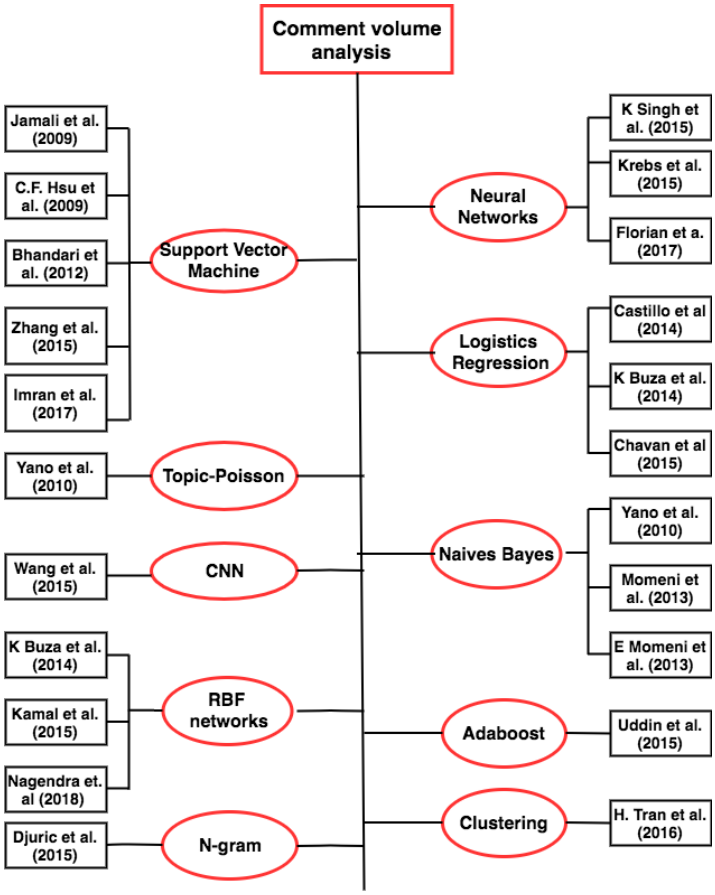


Figure 2.1: Recent techniques used for comment volume prediction

photos) and YouTube (comments on videos). Machine learning algorithms have been used such as logistic regression and naives bayes which are evaluated under the light of precision, recall, F1 measure and ROC. In case of Flickr social data, useful comments have shown more significance which are considered dependant on surface level, entity type and time and yielded precision value 0.87 and recall 0.90. Whereas, YouTube data has yielded precision .65 and 0.83 as recall.

In [69], authors have proposed filtration and classification of useful comments using feature engineering and machine learning techniques on Flickr comments data on 6 collections. Features are extracted from the comments on the basis of author and social features which includes links, trackbacks and network. Semantic and topic features include content based, tone of the subject and polarity based features. Text and Linguistic based features includes readability features, marks and symbols in the comments and information level of the comment. Character-

istics of useful features has been analyzed on 32,132 comments on 11,130 photos. After extracting the significant features training data with total 2000 comments, equal in proportion with useful and un-useful comments. Thus classification is performed using logistics regression, SVM, DT classifier and naives bayes under the light of precision, recall, F-1 measure and area under the Receiver Operator Curve (ROC). This work has yielded predictive accuracy of 89%.

The study on importance of comments is proposed in [44], which shows that comments on news articles reflect its level of interest, significance and order of rank in comparison to other articles. Prediction of comments is performed on seven news agents data of time period of Nov 2008 to April 2009 for calculating the number of comments a news article receive after publication. Correlation measure is used to calculate the relationship between reference of the news article and its publication time. For this, authors have evaluated prediction using statistical methods such as log norm distribution and negative binomial distribution. The experimental results showed that proposed technique gained relative squared error of less than 0.2 for all data sets.

The prediction of the news article and publication classification on Twitter is proposed in [45] using regression and classification algorithms. Feature extraction has been performed on total of 44,000 news article on the basis of score values using correlation in category, subjectivity and source, out of which 2000 article were discarded due to their irrelevancy. Regression algorithms such as logistic regression, SVR have been used for prediction of popularity in terms of comments under the light of evaluation measure r-squared measure. While in case of classification, analysis of news articles has been done that which article will get publication on Twitter with classifier DT, SVM, NB under the light of evaluation measure, accuracy. Results shows that in regression case, logistic regression and SVM have achieved r-squared value of 0.34 & 0.32 respectively. While in classification SVM has accuracy of 81.54%, NB 77.79%, bagging 83.96%, DT 83.75.

Summary of the recent techniques used by many researchers using various techniques is described in Table (2.1). Figure (2.1) depicts the recent techniques used in the comment volume prediction. Table (2.2) provides comparative analysis of

four algorithms which includes Least square regression, Neural networks, Fuzzy system and ANFIS under the evaluation of various important factors, which forms the performance of the algorithm. Here, ✓shows yes for that factor, × no whereas − represents partial answer.

Table 2.2: Comparative analysis of algorithms using various important factors

	Least squares regression	**Neural Networks**	**Fuzzy System**	**ANFIS**
Model free	×	✓	✓	✓
Can resist outliers	×	×	−	−
Explains output	−	×	✓	✓
Suits small dataset	×	×	✓	−
Can be adjusted for new data	×	−	−	−
Reasoning process is visible	✓	×	✓	−
Suits complex models	×	✓	✓	✓
Include known facts	−	−	✓	✓

Chapter 3

Problem Statement

In this chapter, research gaps analyzed in the literature are discussed, problem is defined and contributions of the proposed thesis are also included.

3.1 Research Gaps

Although opinion mining and other techniques retrieving information have shown remarkable work in the area of web, but these common methodologies are not sufficient for blogs. As analyzing the volume of response from blogs counts as a first metric for examining popularity but these days more emphasis has been given on the topics, communities, their opinions and sentiments. Due to timeliness and extremely dynamic network of blogs, predicting volume of response in a quick period of time is an important task. In most of state of art methods, comments are examined on the basis of their extracting features and classifying which is a time consuming task where calculating number of comments is used as a metric for popularity. As there are websites which are only intended to calculate the volume of response, so there is a need of technique which can predict the volume of response a blog will receive. It can be a useful metric for the new bloggers for analyzing their blogs. As per the reviewed literature, various methods are recently used for analysis of blogs such as feature extractions techniques, SVM, Naives Bayes, decision trees and bag of words. These techniques collects features from the comments which can be a time consuming task. So, prediction with ANFIS parameters tuning with stochastic optimization methods can be used which shows very less information redundancy and useful in predicting the volume of response of blogs. These techniques are not used in this area before.

3.2 Statement

To overcome the above limitations, a new technique is proposed for predicting the blog response volume. The new proposed technique is ANFIS with its parameters

tuning with the help of stochastic optimization methods which includes PSO and GA. So this combination is an improved version of ANFIS. In this study blog feedback is selected for predicting the volume of responses a blog will receive after some hours of publishing using ANFIS with PSO and GA. First feedback are split into important features and then PCA is applied to reduce the dimensionality of the data and to keep the only significant information. Then, feature matrix with reduced dimensions are put into the ANFIS algorithm for prediction of volume of response.

3.3 Contribution

- Description of the importance of analysis of blog response volume which can be primary metric for popularity for the users who are intended to count only volume.
- To automate the prediction of responses volume of blog without need of expert using machine learning techniques and proves to be an effective and faster framework.
- Proposed the new users who can start forecasting their blog popularity with this efficient technique. ANFIS combination with stochastic optimization approaches such as PSO and GA results in better performance and reduces training and validation errors.

Chapter 4

Proposed Methodology

Discussing the problem and research gaps of the problem, this section focuses on the methodology proposed in our work. Figure (4.1) depicts the block diagram for prediction process of response volume using ANFIS in combination with Particle Swarm optimization and genetic Algorithm. Firstly, data is loaded and its features are identified. Important features from the data have been combined into feature sets such as basic, basic integration with parent, weekday and textual features. After pre-processing of feature sets, PCA is applied to every feature set for reducing the dimensions and training data is obtained. This data is fed into ANFIS regression model whose parameters are tuned using PSO and GA which are the proposed methods so that it can predict the response volume in blogs. The main goal of using PSO and GA to tune parameters of ANFIS is to reduce the objective function, Mean Squared Error (MSE).

4.1 Adaptive Neuro Fuzzy Inference System

ANFIS was first presented by [79] which is a combination of fuzzy inference system and ANN framework. It has gained popularity in predicting problems of several fields such as stock price [80], pollution [81] and weather [82]. Nodes in ANFIS structure are interconnected where each nodes has its own task according to the incoming signals and parameters present in node. First and fourth nodes are adaptive while others are constant. ANFIS is composed of fuzzy theory and if-then rules. These rules play role while modeling problems. ANFIS with five layers and two rules is depicted in Figure (4.2).

First-order Sugeno fuzzy model [83] represents set of if-then rules described below:
1. if x is A_1 and y is B_1, Then as given in equation (4.1).

$$f_1 = p_1 x + q_1 y + r_1 \tag{4.1}$$

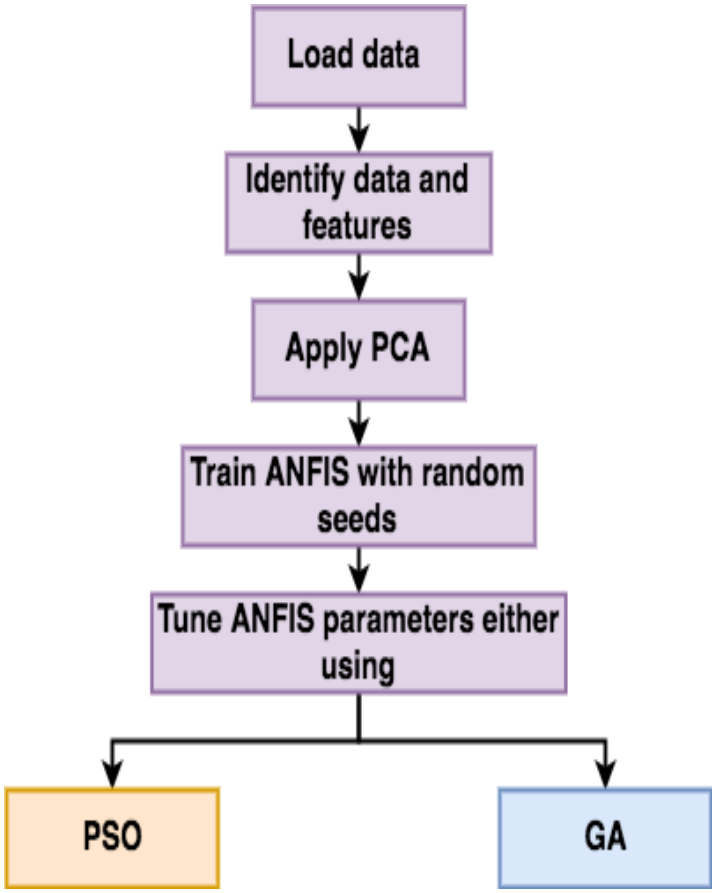


Figure 4.1: Proposed method for prediction

2. if x is A_2 and y is B_2, Then as given in equation (4.2).

$$f_2 = p_2x + q_2y + r_2 \tag{4.2}$$

3. as given in equation (4.3).

$$f = \frac{w_1f_1 + w_2f_2}{w_1 + w_3} \tag{4.3}$$

4.2 Five Layers of ANFIS

ANFIS structure consists of five layers and output of each layer can be stated as follows:

Layer 1: Nodes in this layer are adaptive. Let $O_{l,n}$ is the output of nth node at

layer l.

$$O_{1,1} = \mu_{A_1(x)}, O_{1,2} = \mu_{A_2(x)} O_{1,3} = \mu_{B_1(x)}, O_{1,4} = \mu_{B_2(x)}$$

Here, $O_{1,1}, O_{1,2}, O_{1,3}, O_{1,4}$ refers to the membership values of fuzzy set A_1, A_2, A_3, A_4 and $\mu_A(x) = (1 + |\frac{x-c_i}{a_i}|^{2b_i})^{-1}$ where a_i, b_i, c_i are called as premise parameters.

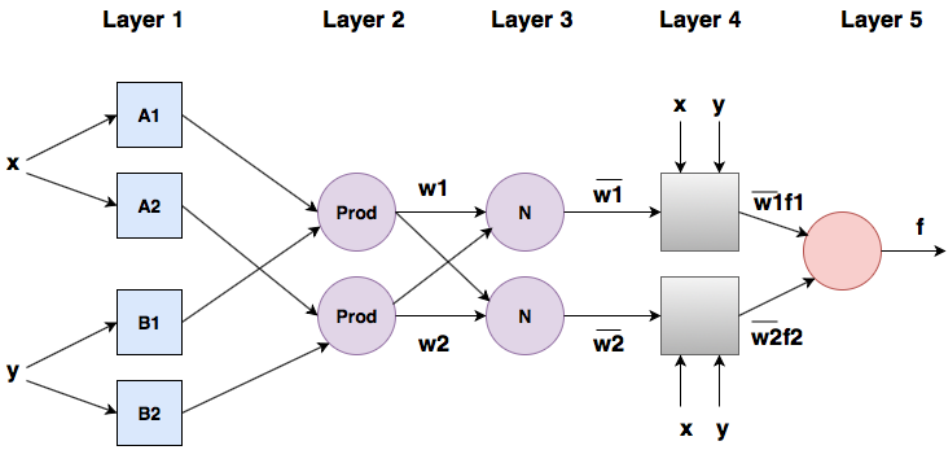


Figure 4.2: Five layers of ANFIS

Layer 2: The value of nodes is constant in this layer. This layer generates the product of all output from previous layer and output of this layer is denoted as follows:

$$O_{2,1} = \mu_{A_1(x)} * \mu_{B_1(y)} = w_1 O_{2,2} = \mu_{A_2(x)} * \mu_{B_2(y)} = w_2$$

Layer 3: The value of node is fixed in this layer and nodes are stated as Norm. Then,

$$O_{3,1} = \frac{w_1}{w_1 + w_2} = \bar{w_1}, O_{3,2} = \frac{w_2}{w_1 + w_2} = \bar{w_2}$$

Layer 4: Nodes in this layer produces each rule output. Let (p_t, q_t, r_t) be a parameter set stated as consequent parameters. So,

$$O_{4,t} = \bar{w_t} * f_t = w_t * (p_t(x) + q_t(x) + r_t), where t = 1, 2.$$

Layer 5: The single node in this layer is a fixed node and generates the final output of all signals:

$$O_{5,t} = \sum_{t} \bar{w}_t f_t, wheret = 1, 2.$$

The premise parameters at layer 1 and consequent parameters present at layer 4 are depicted in Figure (4.3). These parameters are known as the tunable parameters which can be tuned by using optimization methods such as PSO and GA.

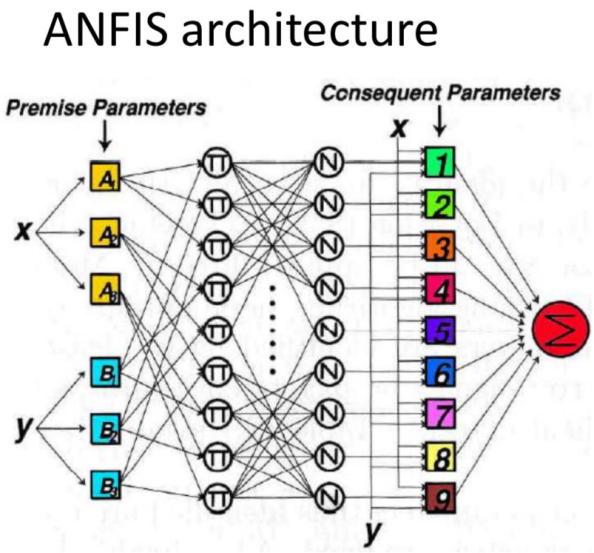


Figure 4.3: Illustration of tunable parameters of ANFIS

4.3 Particle Swarm Optimization

The origin of PSO is inspired by the life of flocking birds and fish schooling [84], [85]. It derives group or swarms of individuals known as particles. In swarm, particles move by learning under social behaviour. PSO search the global best solution by observing the personal best and global best of entire population upon every iteration. Thus, it is known as global optimization algorithm which search for for the solution in n dimensional space [86]. PSO proves better than Ant colony algorithm [87] and EAs [88] as it requires less computational codes, thus inexpensive in terms of memory and speed. It has gained popularity due to simple and effective solution. Particles move in the space towards the location of G_{best} due to social learning in the same group. Every particle has a position and it keeps on varying along with time with every movement. $X_i(t)$ refers to $_ith$ particle

position at time interval t. Moreover, every particle possess velocity to move in the search space. $V_i(t)$ is the velocity of $_ith$ particle at time interval t. Thus, by updating velocity to particle's position, new position of the particle can be derived. The whole process of updating velocity and position vector is depicted in Figure (4.4). The updating velocity of particle which is used to update the position is described in equation (4.4) & (4.5) respectively where w, C_1 and C_2 are the constant variables .

$$V_i(t+1) = w * V_i(t) + C_1(P_i(t) - X_i(t)) + C_2(g(t) - X_i(t)) \tag{4.4}$$

$$X(t+1) = X_i(t) + V_i(t+1) \tag{4.5}$$

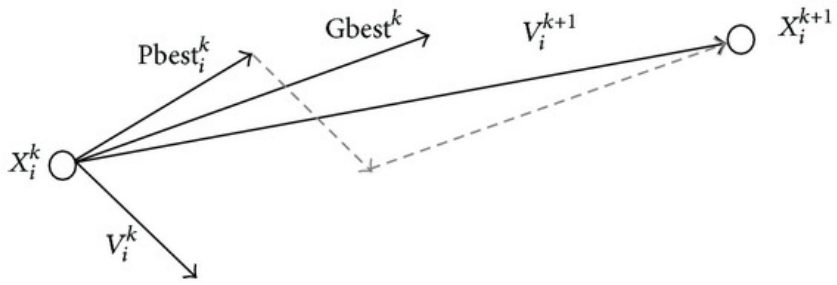


Figure 4.4: Illustration of a particle movement in PSO

Cost function checks for the suitability of the particle position. Particle possess ability to remember their best position for whole of their life. The individual best experience of a particle or best position achieved is known as P_{best}. While best position achieved by whole swarm is known as global best, G_{best}. The particle's experience and social learning is reflected by particle velocity vector. In the search space, every particle possess two components which are described as follows:

- Social component:

  $$y_i(t) - x_i(t)$$

  denotes the best experience gained by a individual particle

- Cognitive component equation:

$$\hat{y}_i(t) - x_i(t)$$

denotes the experience gained by whole group

In PSO algorithm, calculations of velocity vector are noted based on the cognitive and social components. The whole process of PSO algorithm, initializing the random particles then calculating fitness value and evaluating the best solution is described in block diagram in Figure (4.5).

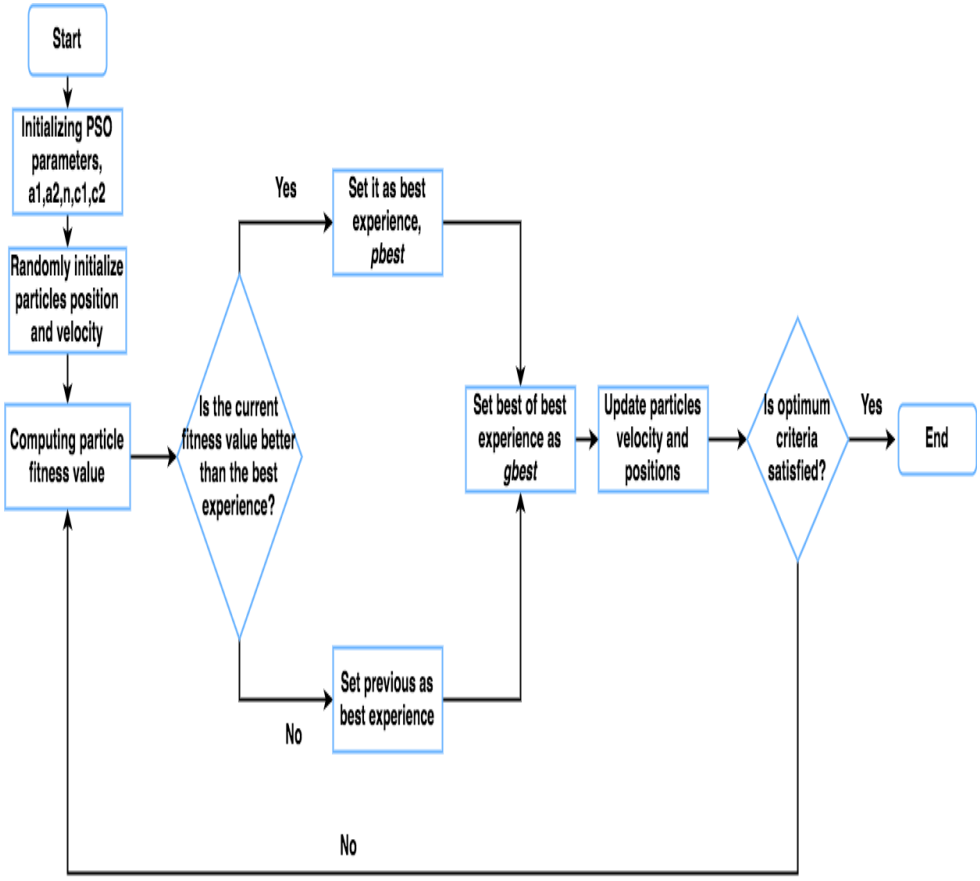


Figure 4.5: Process of PSO technique

4.4 Genetic Algorithm

Genetic Algorithms (GA) are stochastic optimization methods which have the ability to search in wide and complex space. These methods search in the space

by learning from genetics and evolutionary principles [89]. Genetic Algorithms falls into the category of heuristics search. Genetic Algorithm has better abilities than other conventional methods due to the reasons such as it includes various points in search space while eliminating the chance of converging to local optima and considers probabilistic rules instead of deterministic for searching. The generations are formed from randomly initialized individuals, where in every generation fitness of the every individual is calculated. Multiple individuals are selected upon the basis of their fitness value and reproduced to form a new generation. The new generated population is evaluated in the further iteration. Generally, algorithm terminates while satisfying the maximum number of generations or evaluating satisfactory fitness value for the population. GA(s) works in the four main stages such as initialization of random population, selection and reproduction which are illustrated in the Fig (4.6). In the initialization level, initial population is formed by randomly generated individuals where population size is defined by the domain usually covering all the possible solutions. Then, fitness is evaluated of the randomly generated population, then based on the evaluation of fitness value new generation is formed. The first step in generating new population is selection, from each generation some of the portion of population is selected to form new generation. Then, new generation is formed by crossover which is also known as re-combination. Then after selection and cross-over, new generation containing individuals can have copied or re-combined population. To reduce the number of copied individuals, mutation is done by the altering the values. So, with new generated population again repeats the process by evaluating the fitness value [90].

The objective function used for this study, which needs to reduce by combining optimization methods with ANFIS is stated in equation (4.6).

$$MSE = \frac{1}{n}\sum(X_i - X_p)^2 \tag{4.6}$$

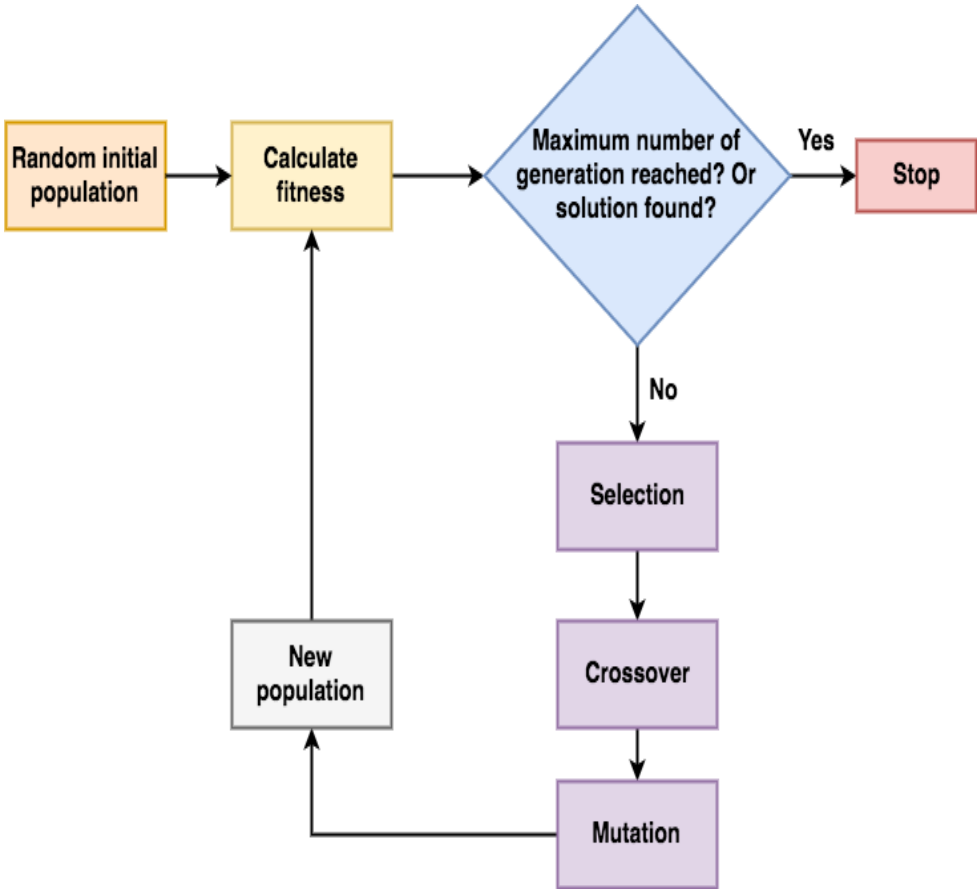


Figure 4.6: Flowchart for genetic algorithm

4.5 Tuning ANFIS using PSO

As PSO algorithm, updates velocity and position of every individual. By updating the position and moving in search space, particle best position experienced by swarm are obtained. Then, for all particles in the space particle best experience and group best experience are generated. This process repeats for all particles until reaching a threshold value which can be on completing the number of iterations or completing the time to determined value of error. The initializing parameters of this proposed are shown in table (4.1). In ANFIS model, there are two types of parameters which can be tuned by PSO, (i) antecedent and (ii) consequent parameters which are explained in section (4.2). Finally, after completing training, best particles from swarm are used to tune ANFIS network parameters. The process of tuning parameters of ANFIS using PSO is depicted in pseudo-code of Algorithm

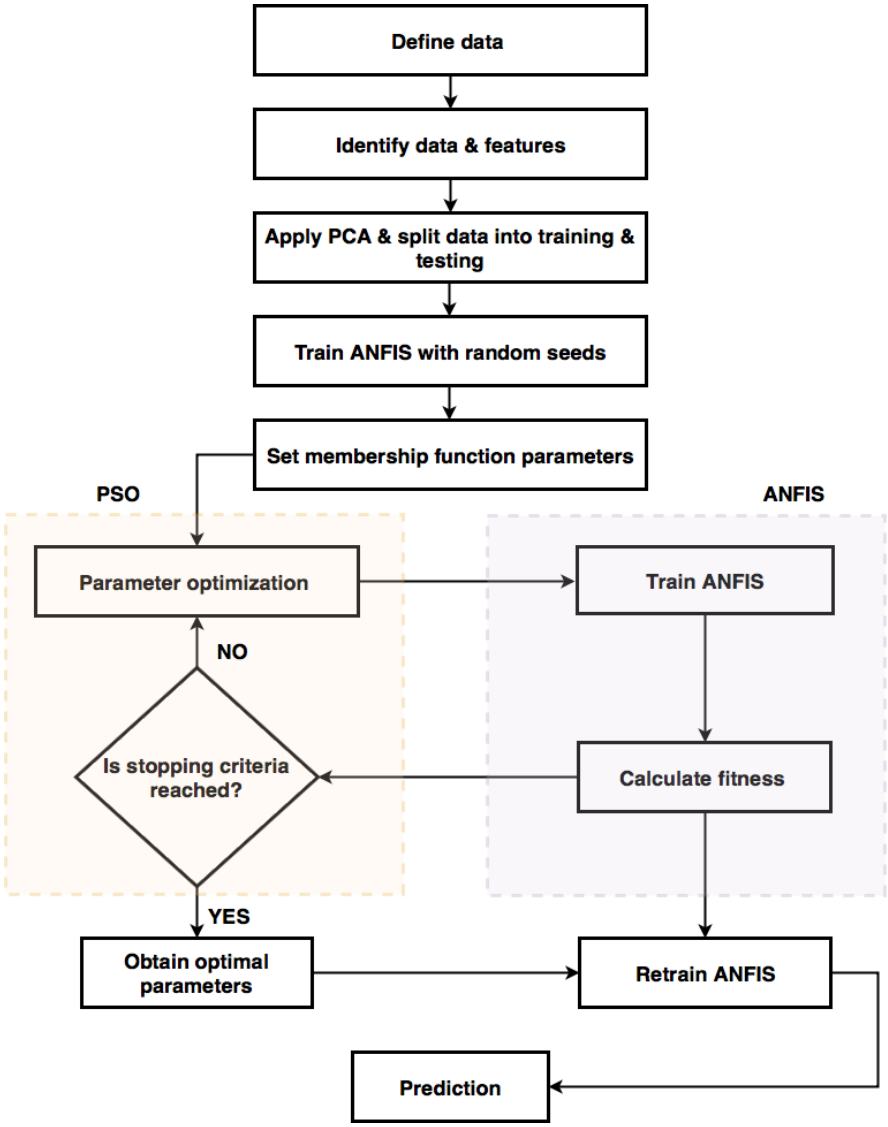


Figure 4.7: Flowchart of proposed technique (i) ANFIS+PSO

(4.1). It will improve the basic ANFIS system training model. The block diagram for this work is illustrated in Fig (4.7).

4.6 Tuning ANFIS using GA

This section explains the method of tuning parameters of ANFIS using GA for prediction of response of blog. Genetic algorithm is an evolutionary method which in combination with ANFIS, optimizes the parameters of the clusters. It is a well known algorithm for its effective and robust suitability over discontinuous and

Algorithm 4.1 Proposed ANFIS-PSO

1: Create the ANFIS network and initialize parameters of ANFIS
2: Call PSO algorithm
3: Evaluate the objective function using equation
4: Repeat step 2 & 3 until best solution is obtained
5: Store the best value of objective function in G_{best}
6: End PSO algorithm
7: Consider the G_{best} value for tuning ANFIS parameters

multi-modal functions [89]. The proposed work of combining ANFIS with GA is illustrated in Fig (4.8) and parameters used in proposed technique are shown in table (4.2). This methods works on two levels which are explained as follows:

- Stochastic optimization involves ANFIS in every generation which evaluates the fitness value of the clustering parameters used. The aim of this algorithm is to reduce the training and cross-validation error of the ANFIS. This level iterates till the stopping condition is fulfilled. If this condition returns yes, then optimal parameters are obtained, otherwise parameters are updated in next generation of GA and whole process is repeated till the search is over.
- After finding the optimal parameters, ANFIS model sets antecedent and consequent parameters. Thus, special training of ANFIS is performed by tuning parameters, then prediction performance is evaluated by feeding the testing data. The pseudo-code in Algorithm (4.2) explains the tuning process of ANFIS using GA.

Algorithm 4.2 Proposed ANFIS-GA

1: Create the ANFIS network and initialize the parameters of GA i.e. Pop size, selection and mutation
2: Evaluate fitness function
3: **if** minimum population reached **then**
4: Stop
5: **else**
6: Evaluate the selection, and reproduction phase
7: **end if**
8: Generate the new population & evaluate the fitness function using equation (4.6)
9: Repeat the steps 3 to 4 until the best fitness value is obtained
10: Set the best fitness value for tuning ANFIS parameters

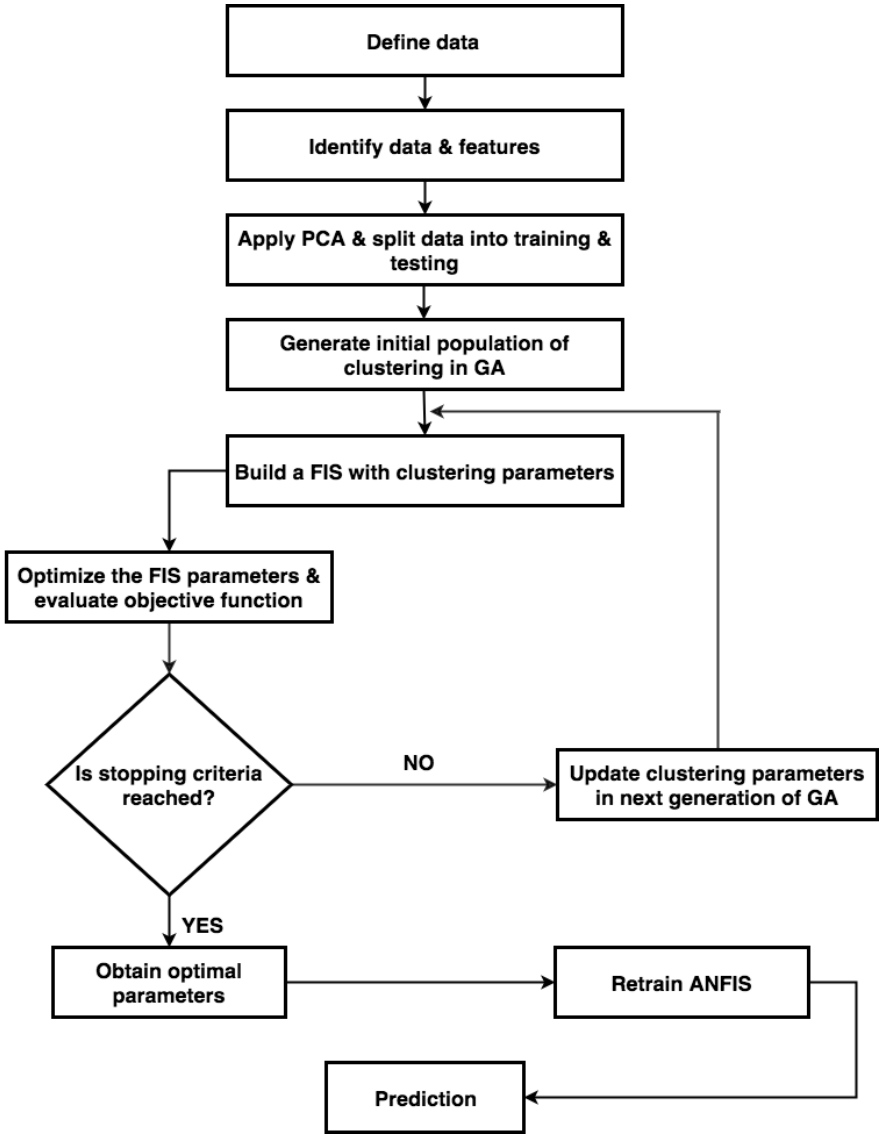


Figure 4.8: Proposed technique (ii) ANFIS+GA

Table 4.1: Parameters of ANFIS+PSO algorithm

Maximum Iterations of PSO	1000
Number of fuzzy sets for each input	3
Size of population	25
Inertia weight	1
Individual learning coefficient, C_1	1
Group learning coefficient, C_2	2
Maximum particle velocity	$V_{max} = 1.2$

Table 4.2: Parameters of ANFIS+GA algorithm

pc (crossover%)	0.4
pm (mutation%)	0.7
gamma	0.7
mutation rate	0.15
beta	8

Chapter 5

Experimental Results

5.1 Dataset Description

Documents from Hungarian blog sites are used as dataset. This is public data available online on UCI online repository. The origin of the dataset is fully described in [58]. There are about total 280 attributes in data except one target variable and 52397 entries in the data. Through changing the document data into vectors for applying regression models, data has been formed into following features by extracting them each document. Dataset is split on basic, textual, weekday and parent features on the basis of importance of domain concepts for analysis. The division of data into feature sets is illustrated in Figure (5.1).

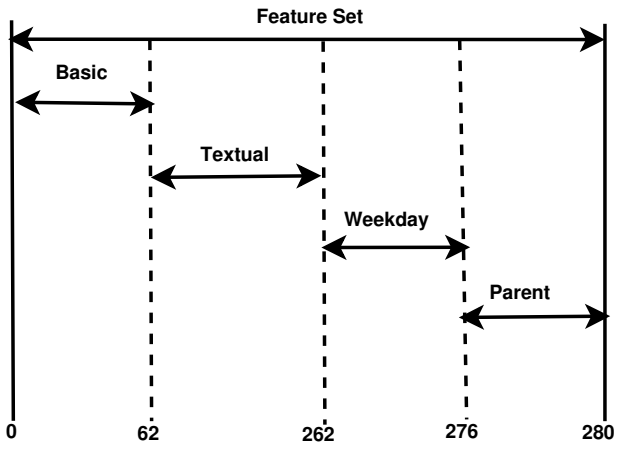


Figure 5.1: Division of feature sets using Feature extraction

1. *Basic Features*: it reflects the importance of links and feedback of documents. In relation to baseTime, it denotes the count of links and responses a document receives. Count of links and responses in a time period from 48 hrs to 24 hrs advance to baseTime. It represents the variations in the count of links and responses in relation to baseTime. Amount of links and responses after 24 hrs of posting. The sum total of all above fields.

2. *Textual Features*: it reflects the importance of discriminative words. The amount of most critical words.
3. *Weekday Features*: it shows the importance of decision of days for publication and prediction. It lies between [0,1], thus binary indicator which represents the day of week document published and which day forecasting is to be done.
4. *Parent Features*: it reflects the importance of trackback links between text documents, count of parents. Moreover, minimum, maximum and average of responses that parents obtained.
Here baseTime refers to selected date and time, as in our study documents published in 72 hrs prior to basetime are considered as new and prediction is performed. As older documents from this generally stops receiving new comments.

5.2 Principal Component Analysis

For improving the performance of the predictive algorithms, analysis of significant features from the data is imported. As there are many dimensionality reducing algorithms such as PCA [91], Independent Component Analysis (ICA) [92], Linear Discriminant Analysis (LDA) [93] and Generalized Discriminant (GDA) [94]. In this study, to avoid the curse of dimensionality and faster execution, PCA which is an unsupervised learning algorithm is used because it lose the least information while reducing the data and converting into new variables. Moreover, it can fit with the normal data and derives features in a orthogonal projection. Figure (5.2) provides the ratio of information of the first five principal components. The illustration of information gain is depicted in Figure (5.3). After extracting important information and to reduce the dimensionality of data using PCA from the features sets, data is split into subsets, training (70%, 36,677) and testing set (30%, 15720). The analysis of information gain from each of the feature set is shown in table (5.1).

5.3 Evaluation metrics

To evaluate the ANFIS combination with PSO and GA based regression model, mean-squared error which is objective function, accuracy and correlation (r) is

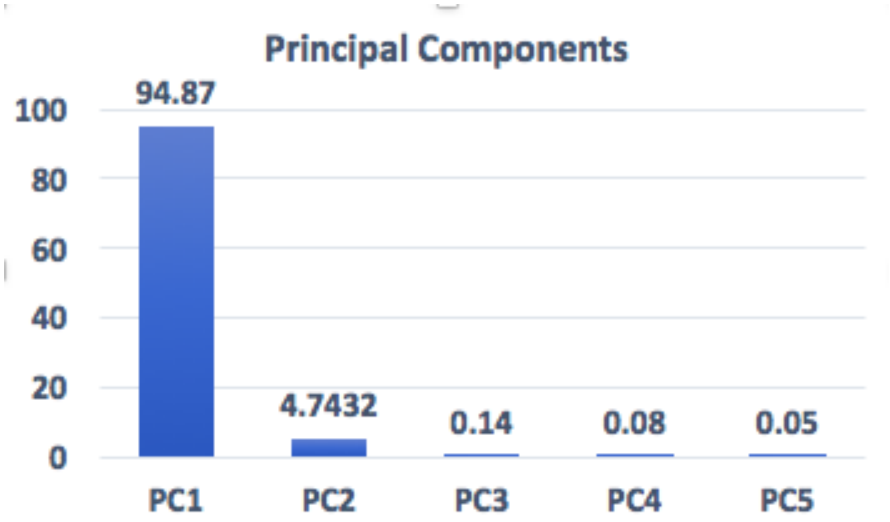


Figure 5.2: Information gain for the principal components

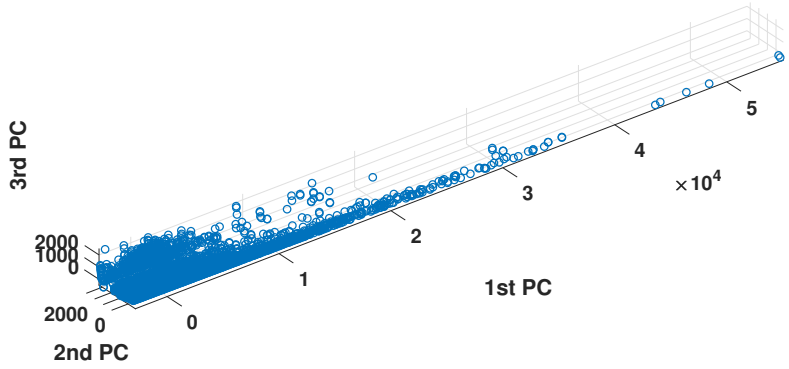


Figure 5.3: Illustration of data variability of Principal Components

used for examining the results and reliability of blog response volume prediction, which are described as follows:

$$MSE = \frac{1}{n}\sum(X_i - X_p)^2$$

$$r = \frac{\frac{\Sigma_i(X_i - X_p)^2}{n-l-1}}{\frac{\Sigma_i(X_i - \bar{X})^2}{n-1}}$$

Table 5.1: Information gain per PCA components using different sets

PC1	94.89	94.88	94.87	94.89	94.87
PC2	4.74	4.74	4.74	4.74	4.732
PC3	0.14	0.14	0.14	0.14	0.14
PC4	0.08	0.08	0.08	0.08	0.08
PC5	0.04	0.04	0.05	0.04	0.05

$$Accuracy = \frac{\sum_i if(X_i - X_p| \leq e_r)}{n}$$

where X_i is denotes the actual values and X_p denotes the predicted values using the regression algorithm, $\bar{X}$ is the average value of n samples. While l denotes various independent attributes and e_r is the acceptance of error rate.

After getting the training data reduced dimensionality, results are obtained evaluating the performance of regression models which includes SVM, ANN, ANFIS, ANFIS+PSO, ANFIS+GA on different feature sets under the light of evaluation metrics correlation, mean-squared error and accuracy. Computational time of training and testing of each model is also considered which is presented in seconds. Table (5.2) represents the results using information gained by first principal component. First principal component analysis marked huge information gain, as to compare and evaluate the performance results are also obtained using first and second principal components which are shown in Table (5.3).

Time taken by the regression models includes time for training and evaluate the testing data. This evaluation measure is depicted in graph in Figure (5.4) which shows the time taken by using data gained by Ist principal component on different feature sets which are Basic(B), Basic+Weekday(B+W), Basic+Parent(B+P), Basic+Textual(B+T) and all features. The graphs shows that ANFIS+GA has the maximum computational time as compared to other algorithms over all feature sets. Whereas, neural networks took the least time for whole process of prediction. Using first and second principal component time taken on different feature sets is represented in the graph in Figure (5.5) which shows that using two principal components increased the time taken by algorithms as compared to using information by only one principal component.

Table 5.2: Results with First PC component

		Basic	**B+W**	**B+P**	**B+T**	**All**
SVM	r	0.65	0.67	0.67	0.67	0.68
	MSE	1550.40	1429.10	1453.20	1501.00	1402.30
	Acc.(%)	84.35	84.70	84.40	84.30	85.20
	Time	72	76	73	76	75
ANN	r	0.74	0.75	0.75	0.76	0.77
	MSE	1470	1422	1428	1398	1369
	Acc.(%)	86.11	87	86	87	87
	Time	30	32	34	29	34
ANFIS	r	0.79	0.8	0.79	0.8	0.81
	MSE	1424.38	1435	1401	1376	1311
	Acc.(%)	89	89	89	89	90
	Time	102	100	103	99	104
+GA	r	0.81	0.82	0.8	0.82	0.83
	MSE	1103	1090	1077	1053	1008
	Acc.(%)	89	90	90	90	91
	Time	188	185	182	186	190
+PSO	r	0.82	0.83	0.82	0.83	0.84
	MSE	1093.24	988	980	1001	926
	Acc.(%)	90.02	91	90.00	91	92
	Time	163	164	166	162	166

Table 5.3: Performance metrics with first two principal components

		Basic	**B+W**	**B+P**	**B+T**	**All**
SVM	r	0.82	0.84	0.82	0.82	0.85
	MSE	1478.5	1471.34	1474.6	1476	1470.4
	Acc.(%)	86	89	86	89	91
	Time	84	85	89	87	89
ANN	r	0.87	0.89	0.88	0.89	0.9
	MSE	1383	1381	1378	1376	1370
	Acc.(%)	87	88	88	90	91
	Time	59	61	65	63	65
ANFIS	r	0.92	0.93	0.94	0.93	0.94
	MSE	1244.1	1243.4	1242.2	1241.8	1239.7
	Acc.(%)	91	92	93	93	94
	Time	134	135	137	138	140
+GA	r	0.94	0.95	0.95	0.95	0.96
	MSE	1008.12	1006.4	1007.2	1004.2	1003.6
	Acc.(%)	94	94.2	94.2	94.6	94.60
	Time	237	238	239	238	240
+PSO	r	0.96	0.96	0.97	0.96	0.97
	MSE	1005.8	1004.6	1004.2	1003.4	1003.2
	Acc.(%)	94.85	94.91	94.92	94.91	95
	Time	220	223	224	224	225

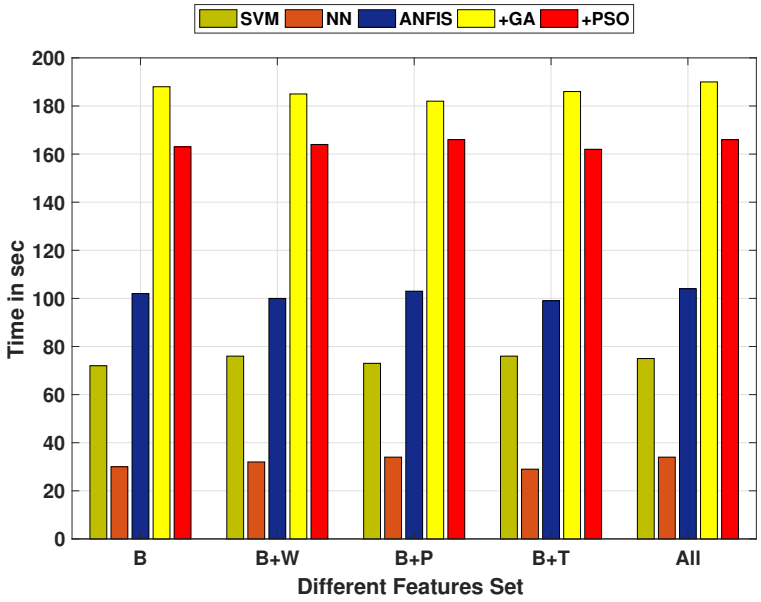


Figure 5.4: Time taken by 1st PC on different feature set

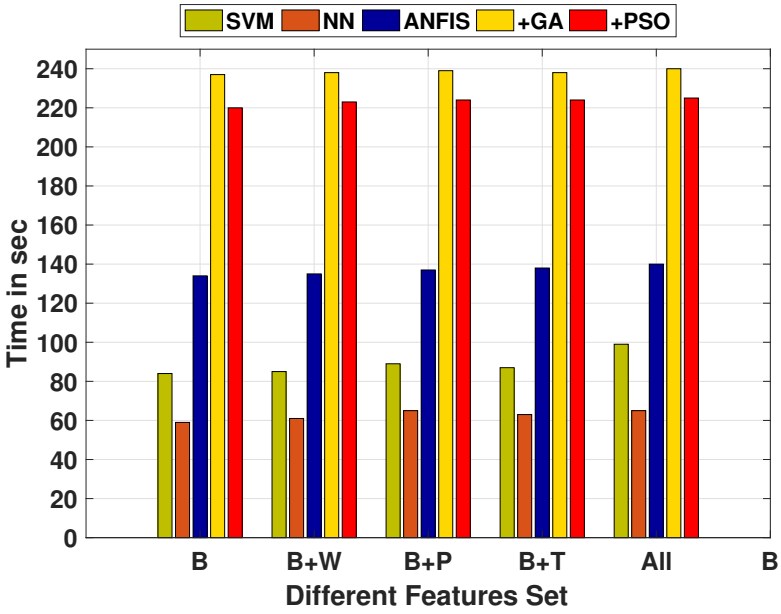


Figure 5.5: Time taken by two principal components on feature set

5.3.1 Accuracy

In the prediction problems, accuracy acts as a major evaluation metric for checking the predictive performance of the algorithm. Accuracy using first principal

component is depicted in graph in Figure (5.6) which shows that ANFIS+PSO gains the maximum accuracy on every feature used, ANFIS+GA also shows comparative accuracy. In this scenario, SVM has gained the least accuracy.

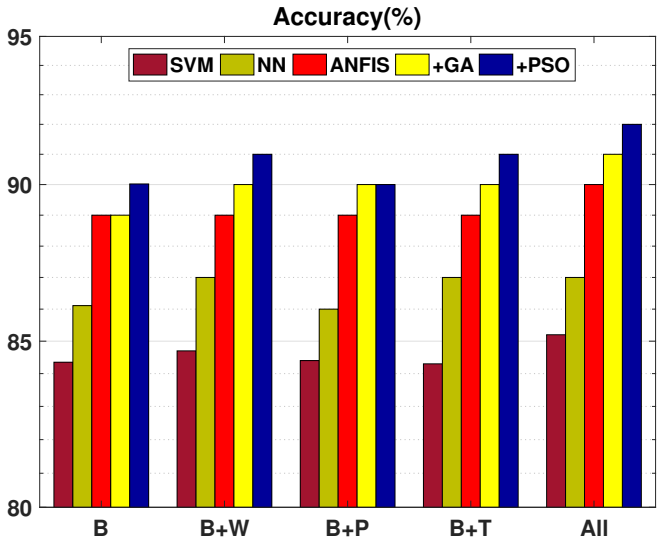


Figure 5.6: Accuracy using first PC

Using two principal components, accuracy is depicted in graph in Figure (5.7) which shows comparison of results of all algorithms. It shows that ANFIS+PSO has gained maximum accuracy on each of the feature set while in this scenario ANFIS+GA has the most comparative results to it.

5.3.2 Correlation

Comparison of correlation is analyzed using first principal component and first two principal components. Using first principal components, ANFIS+PSO gained maximum correlation value of 0.84 whereas ANFIS + GA has gained the same correlation value as 0.83 which is depicted in graph in Figure (5.8). In the graph in Figure (5.9) using first two components, ANFIS+PSO has gained 0.97 correlation value, ANFIS+GA has scored 0.96, ANFIS has 0.94, ANN has 0.9 and SVM 0.85.

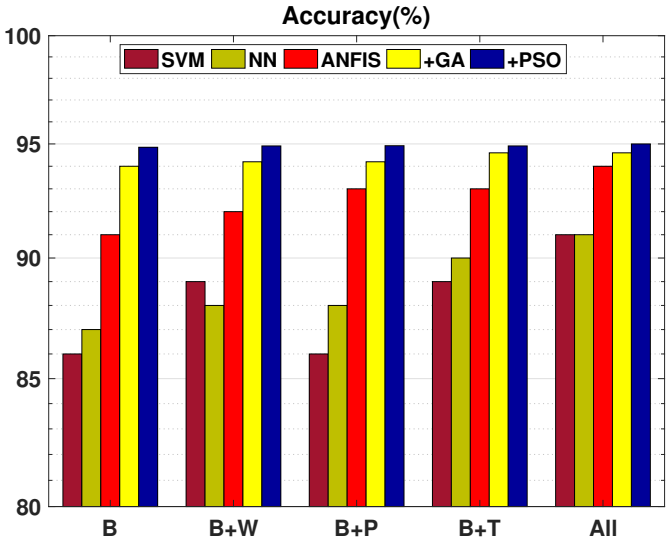


Figure 5.7: Accuracy using two principal Components

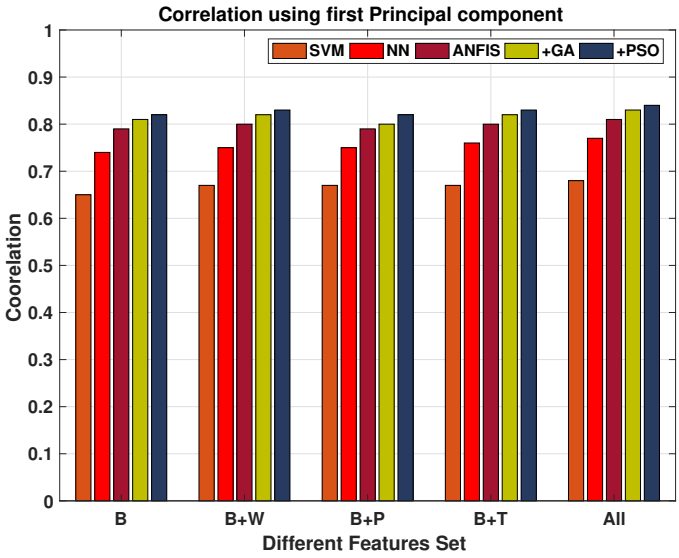


Figure 5.8: Correlation using first principal component

5.3.3 Mean Squared Error

Mean squared error gained by all algorithms on all data. In this, SVM has gained the maximum value of error, 1470.04. Whereas ANFIS+PSO has the minimum value of 987.64.

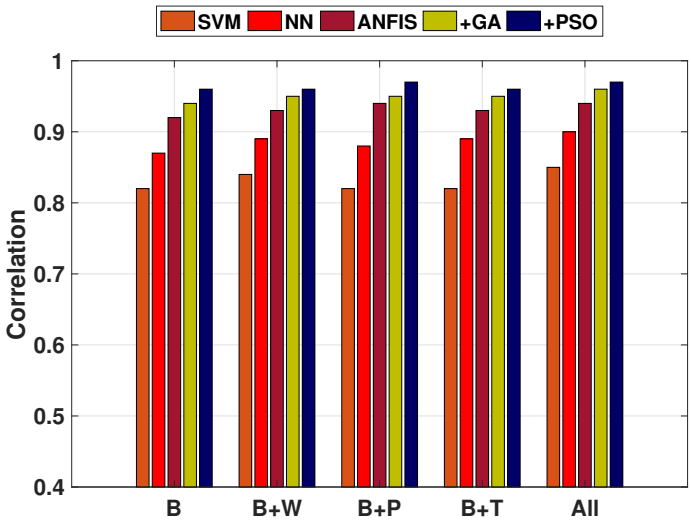
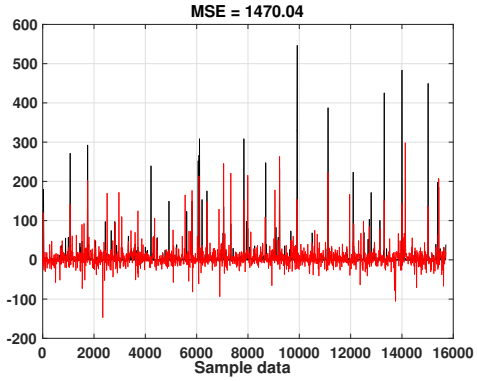
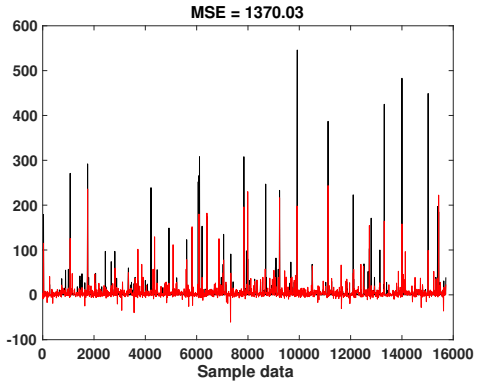


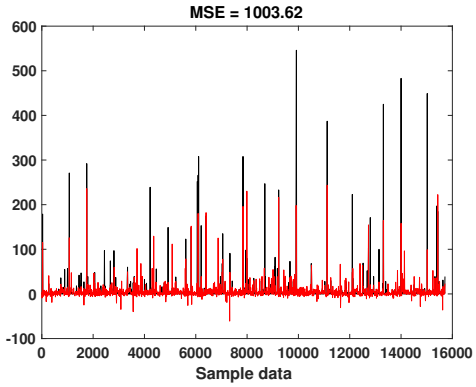
Figure 5.9: Correlation using two principal components



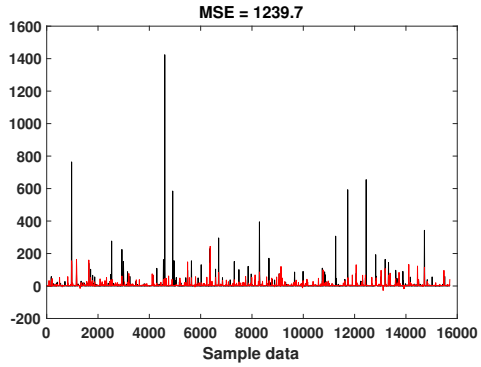
(a) Mean squared error value of SVM



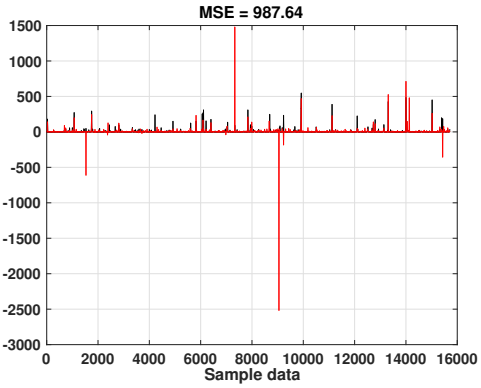
(b) Mean squared error value of NN



(c) Mean squared error value of ANFIS



(d) Mean squared error value of ANFIS+GA



(e) Mean squared error value of ANFIS+PSO

Figure 5.10: MSE of algorithms using All data

Chapter 6

Conclusion and Future Work

In this thesis, a hybrid ANFIS is proposed which combines with stochastic optimization methods for parameters tuning. From the data, important features which are basic features and its combination with other features such as parent, textual and weekend features have been derived. PCA is applied on each of the feature set to reduce the dimensionality while maintaining only significant information. Then data is fed into proposed Adaptive Neuro Fuzzy Inference System which in combination of PSO and GA for prediction of blog response volume. After experimentation results are obtained under the light of evaluation metrics such as accuracy, mean squared error and correlation which shows that stochastic optimization methods used with ANFIS outperformed the conventional methods which includes support vector machines, ANFIS and neural networks because these methods search by implementing statistical analysis and probabilistic rules. However, analyzing the stochastic optimization methods it is found that PSO shows better results than GA overall and takes less computational time as compared to GA because of its expensive computational time and complex structure.

Limitations of this research that it is a difficult task to analyze all features from the document for prediction, only the most important features should be used which can make prediction even faster. As, in this research basic features has the most importance. So, analysis of importance of features is important

Future addition to this thesis could be to use the posts data from real world applications, Twitter and Facebook. Other feature extraction techniques, parameters tuning and sentiment analysis techniques can be incorporated. Moreover, this work can be extended to analyze the attitude of the people from the response submitted.

References

[1] M. Sterling, P. Leung, D. Wright, and T. F. Bishop, "The use of social media in graduate medical education: a systematic review," *Academic Medicine*, vol. 92, no. 7, pp. 1043–1056, 2017.

[2] J. Arafat, M. A. Habib, and R. Hossain, "Analyzing public emotion and predicting stock market using social media," *American Journal of Engineering Research*, vol. 2, no. 9, pp. 265–75, 2013.

[3] E. Bonsón, S. Royo, and M. Ratkai, "Citizens' engagement on local governments' facebook sites. an empirical analysis: The impact of different media and content types in western europe," *Government Information Quarterly*, vol. 32, no. 1, pp. 52–62, 2015.

[4] M. Taylor, "Step-by-step wordpress for beginners: How to build a beautiful website on your own domain from scratch (video course included)," 2016.

[5] J. Greer and P.-L. Pan, "The role of website format, blog use, and information-gathering acquaintance in online message assessment," *Telematics and Informatics*, vol. 32, no. 4, pp. 594–602, 2015.

[6] O. Serrat, "Social media and the public sector," in *Knowledge Solutions*. Springer, 2017, pp. 925–935.

[7] G. Mishne, N. Glance *et al.*, "Leave a reply: An analysis of weblog comments," in *Third annual workshop on the Weblogging ecosystem*. Edinburgh, Scotland, 2006.

[8] J. Kim, U. Yun, G. Pyun, H. Ryang, G. Lee, E. Yoon, and K. H. Ryu, "A blog ranking algorithm using analysis of both blog influence and characteristics of blog posts," *Cluster Computing*, vol. 18, no. 1, pp. 157–164, 2015.

[9] M. D. F. Emerson, N. M. Pereira, J. L. Sardenberg, K. A. Albuquerque, M. C. Costa, C. L. Seifert, M. H. C. Ribeiro, and J. B. D. M. Rios, "Allergy blog:

health information," in *World Allergy Organization Journal*, vol. 8, no. 1. BioMed Central, 2015, p. A136.

[10] D. Ryfe, D. Mensing, and R. Kelley, "What is the meaning of a news link?" *Digital Journalism*, vol. 4, no. 1, pp. 41–54, 2016.

[11] WorldWideWebSize.com, "The size of the world wide web," Feb. 26 2018, national Vulnerability Database.

[12] D. A. Huffaker and S. L. Calvert, "Gender, identity, and language use in teenage blogs," *Journal of computer-mediated communication*, vol. 10, no. 2, p. JCMC10211, 2005.

[13] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Icwsm*, vol. 10, no. 1, pp. 178–185, 2010.

[14] J. Markoff, "Entrepreneurs see a web guided by common sense," *New York Times*, vol. 12, p. 2006, 2006.

[15] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red opal: product-feature scoring from reviews," in *Proceedings of the 8th ACM conference on Electronic commerce*. ACM, 2007, pp. 182–191.

[16] X. Ding, B. Liu, and P. S. Yu, "A holistic lexicon-based approach to opinion mining," in *Proceedings of the 2008 international conference on web search and data mining*. ACM, 2008, pp. 231–240.

[17] S. Kasthuri, L. Jayasimman, and A. N. Jebaseeli, "An opinion mining and sentiment analysis techniques: A survey," *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 2, pp. 2395–0056, 2016.

[18] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[19] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in *LREc*, vol. 10, no. 2010, 2010.

[20] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[21] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[22] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[23] V. K. Singh, R. Piryani, A. Uddin, and P. Waila, "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification," in *Automation, computing, communication, control and compressed sensing (iMac4s), 2013 international multi-conference on*. IEEE, 2013, pp. 712–717.

[24] A. Christin, "Counting clicks: Quantification and variation in web journalism in the united states and france," *American Journal of Sociology*, vol. 123, no. 5, pp. 1382–1415, 2018.

[25] M. Chau and J. Xu, "Business intelligence in blogs: Understanding consumer interactions and communities," *MIS quarterly*, pp. 1189–1216, 2012.

[26] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of computational science*, vol. 2, no. 1, pp. 1–8, 2011.

[27] E. Qualman, *Socialnomics: How social media transforms the way we live and do business*. John Wiley & Sons, 2010.

[28] X. Wang, C. Yu, and Y. Wei, "Social media peer communication and impacts on purchase intentions: A consumer socialization framework," *Journal of interactive marketing*, vol. 26, no. 4, pp. 198–208, 2012.

[29] J. Leskovec, "Social media analytics: tracking, modeling and predicting the flow of information through networks," in *Proceedings of the 20th international conference companion on World wide web*. ACM, 2011, pp. 277–278.

[30] M. J. Kushin and M. Yamamoto, "Did social media really matter? college students' use of online media and political decision making in the 2008 election," *Mass Communication and Society*, vol. 13, no. 5, pp. 608–630, 2010.

[31] I. Persing and V. Ng, "Vote prediction on comments in social polls," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1127–1138.

[32] Q. You, L. Cao, Y. Cong, X. Zhang, and J. Luo, "A multifaceted approach to social multimedia-based prediction of elections," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2271–2280, 2015.

[33] A. Bermingham and A. Smeaton, "On using twitter to monitor political sentiment and predict election results," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, 2011, pp. 2–10.

[34] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media." *ICWSM*, vol. 13, pp. 1–10, 2013.

[35] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 2013, pp. 47–56.

[36] A. Sadilek, H. A. Kautz, and V. Silenzio, "Predicting disease transmission from geo-tagged micro-blog data." in *AAAI*, 2012, pp. 136–142.

[37] S. Chun, S. Shulman, R. Sandoval, and E. Hovy, "Government 2.0: Making connections between citizens, data and government," *Information Polity*, vol. 15, no. 1, 2, pp. 1–9, 2010.

[38] D. Linders, "From e-government to we-government: Defining a typology for citizen coproduction in the age of social media," *Government Information Quarterly*, vol. 29, no. 4, pp. 446–454, 2012.

[39] D. Coursey and D. F. Norris, "Models of e-government: Are they correct? an empirical assessment," *Public administration review*, vol. 68, no. 3, pp. 523–536, 2008.

[40] I. Mergel and S. I. Bretschneider, "A three-stage adoption process for social media use in government," *Public Administration Review*, vol. 73, no. 3, pp. 390–400, 2013.

[41] T. Yano and N. A. Smith, "What's worthy of comment? content and comment volume in political blogs." in *ICWSM*, 2010.

[42] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the crowdsourcing power of social media for disaster relief," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 10–14, 2011.

[43] S. Maine, R. Shute, and G. Martin, "Educating parents about youth suicide: Knowledge, response to suicidal statements, attitudes, and intention to help," *Suicide and Life-Threatening Behavior*, vol. 31, no. 3, pp. 320–332, 2001.

[44] M. Tsagkias, W. Weerkamp, and M. De Rijke, "News comments: Exploring, modeling, and online prediction," in *European Conference on Information Retrieval*. Springer, 2010, pp. 191–203.

[45] R. Bandari, S. Asur, and B. A. Huberman, "The pulse of news in social media: Forecasting popularity," *ICWSM*, vol. 12, pp. 26–33, 2012.

[46] C. Castillo, M. El-Haddad, J. Pfeffer, and M. Stempeck, "Characterizing the life cycle of online news stories using social media reactions," in *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 2014, pp. 211–223.

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

[48] A. Natarajan, B. V. Srinivasan, V. Gupta, A. Ganesan, A. Jain, S. Revankar, J. Singh, and B. Polineni, "Method and apparatus for prediction of community reaction to a post," Dec. 26 2017, uS Patent 9,852,239.

[49] S.-D. Kim, S.-H. Kim, and H.-G. Cho, "Predicting the virtual temperature of web-blog articles as a measurement tool for online popularity," in *Computer*

and Information Technology (CIT), 2011 IEEE 11th International Conference on. IEEE, 2011, pp. 449–454.

[50] I. S. Pantelidis, "Electronic meal experience: A content analysis of online restaurant comments," *Cornell Hospitality Quarterly*, vol. 51, no. 4, pp. 483–491, 2010.

[51] C. M. Cheung, M. K. Lee, and N. Rabjohn, "The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities," *Internet research*, vol. 18, no. 3, pp. 229–247, 2008.

[52] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert systems with applications*, vol. 36, no. 3, pp. 6527–6535, 2009.

[53] J. Miguéns, R. Baggio, and C. Costa, "Social media and tourism destinations: Tripadvisor case study," *Advances in tourism research*, vol. 26, no. 28, pp. 1–6, 2008.

[54] S. A. Wood, A. D. Guerry, J. M. Silver, and M. Lacayo, "Using social media to quantify nature-based tourism and recreation," *Scientific reports*, vol. 3, p. 2976, 2013.

[55] N. Dabbagh and A. Kitsantas, "Personal learning environments, social media, and self-regulated learning: A natural formula for connecting formal and informal learning," *The Internet and higher education*, vol. 15, no. 1, pp. 3–8, 2012.

[56] P. A. Tess, "The role of social media in higher education classes (real and virtual)–a literature review," *Computers in Human Behavior*, vol. 29, no. 5, pp. A60–A68, 2013.

[57] P. Pollara and J. Zhu, "Social networking and education: Using facebook as an edusocial space," in *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 2011, pp. 3330–3338.

[58] K. Buza, "Feedback prediction for blogs," in *Data analysis, machine learning and knowledge discovery*. Springer, 2014, pp. 145–152.

[59] B. Bi and J. Cho, "Modeling a retweet network via an adaptive bayesian approach," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 459–469.

[60] J. Tang, S. Chang, C. Aggarwal, and H. Liu, "Negative link prediction in social media," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 87–96.

[61] H. Tran and M. Shcherbakov, "Detection and prediction of users attitude based on real-time and batch sentiment analysis of facebook comments," in *International Conference on Computational Social Networks*. Springer, 2016, pp. 273–284.

[62] P. Schultes, V. Dorner, and F. Lehner, "Leave a comment! an in-depth analysis of user comments on youtube." *Wirtschaftsinformatik*, vol. 42, pp. 659–673, 2013.

[63] F. Krebs, B. Lubascher, T. Moers, P. Schaap, and G. Spanakis, "Social emotion mining techniques for facebook posts reaction prediction," *arXiv preprint arXiv:1712.03249*, 2017.

[64] M. T. Uddin, "Automated blog feedback prediction with ada-boost classifier," in *Informatics, Electronics & Vision (ICIEV), 2015 International Conference on*. IEEE, 2015, pp. 1–5.

[65] K. Singh, R. Sandhu, and D. Kumar, "Comment volume prediction using neural networks and decision trees," in *IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation, UKSim2015 (UKSim2015), Cambridge, United Kingdom*, 2015.

[66] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Advances in*

computing, communications and informatics (ICACCI), 2015 International Conference on. IEEE, 2015, pp. 2354–2358.

[67] A. Bessi, F. Zollo, M. Del Vicario, M. Puliga, A. Scala, G. Caldarelli, B. Uzzi, and W. Quattrociocchi, "Users polarization on facebook and youtube," *PloS one*, vol. 11, no. 8, p. e0159641, 2016.

[68] J. Wang, A. A. Nabi, G. Wang, C. Wan, and T.-T. Ng, "Towards predicting the likeability of fashion images," *arXiv preprint arXiv:1511.05296*, 2015.

[69] E. Momeni, K. Tao, B. Haslhofer, and G.-J. Houben, "Identification of useful user comments in social media: a case study on flickr commons," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 1–10.

[70] E. Momeni, C. Cardie, and M. Ott, "Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects." in *ICWSM*, 2013.

[71] Y. Artzi, P. Pantel, and M. Gamon, "Predicting responses to microblog posts," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2012, pp. 602–606.

[72] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and svmperf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[73] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao, and S.-F. Chang, "Predicting viewer affective comments based on image content in social media," in *Proceedings of International Conference on Multimedia Retrieval*. ACM, 2014, p. 233.

[74] C.-F. Hsu, E. Khabiri, and J. Caverlee, "Ranking comments on the social web," in *Computational Science and Engineering, 2009. CSE'09. International Conference on*, vol. 4. IEEE, 2009, pp. 90–97.

[75] J. Du, H. Xu, and X. Huang, "Box office prediction based on microblog," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1680–1689, 2014.

[76] S. Boya and M. Singh, "Arousal prediction of news articles in social media," in *Mining Intelligence and Knowledge Exploration: 5th International Conference, MIKE 2017, Hyderabad, India, December 13–15, 2017, Proceedings*, vol. 10682. Springer, 2018, p. 308.

[77] S. Gottipati and J. Jiang, "Finding thoughtful comments from social media," *Proceedings of COLING 2012*, pp. 995–1010, 2012.

[78] A. Imran, W. Aslam, and M. Ullah, "Quantitative prediction of offensiveness using text mining of twitter data," *Sindh University Research Journal-SURJ (Science Series)*, vol. 49, no. 1, 2017.

[79] J.-S. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.

[80] M. Billah, S. Waheed, and A. Hanifa, "Predicting closing stock price using artificial neural network and adaptive neuro fuzzy inference system (anfis): The case of the dhaka stock exchange," *International Journal of Computer Applications*, vol. 129, no. 11, pp. 1–5, 2015.

[81] K. Prasad, A. K. Gorai, and P. Goyal, "Development of anfis models for air quality forecasting and input optimization for reducing the computational cost and time," *Atmospheric environment*, vol. 128, pp. 246–262, 2016.

[82] R. Rajkumar, A. J. Albert, and D. Chandrakala, "A new approach to adaptive neuro-fuzzy modeling using kernel nonnegative matrix factorization (knmf) clustering for weather forecasting," 2017.

[83] U. Kaymak, "An enhanced approach to rule base simplification of first-order takagi-sugeno fuzzy inference systems," in *Advances in Fuzzy Logic and Technology 2017: Proceedings of: EUSFLAT-2017–The 10th Conference of the European Society for Fuzzy Logic and Technology, September 11-15, 2017, Warsaw, Poland IWIFSGN2017–The Sixteenth International Workshop on*

Intuitionistic Fuzzy Sets and Generalized Nets, September 13-15, 2017, Warsaw, Poland, vol. 2. Springer, 2017, p. 92.

[84] S. Garnier, J. Gautrais, and G. Theraulaz, "The biological principles of swarm intelligence," *Swarm Intelligence*, vol. 1, no. 1, pp. 3–31, 2007.

[85] C. Blum and X. Li, "Swarm intelligence in optimization," in *Swarm Intelligence*. Springer, 2008, pp. 43–85.

[86] M. Dorigo, "Ten years of swarm intelligence," *Swarm Intelligence*, vol. 10, no. 4, pp. 245–246, 2016.

[87] K.-L. Du and M. Swamy, "Ant colony optimization," in *Search and Optimization by Metaheuristics*. Springer, 2016, pp. 191–199.

[88] C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds, "An evolutionary algorithm approach to link prediction in dynamic social networks," *Journal of Computational Science*, vol. 5, no. 5, pp. 750–764, 2014.

[89] K.-S. Shin and Y.-J. Lee, "A genetic algorithm application in bankruptcy prediction modeling," *Expert Systems with Applications*, vol. 23, no. 3, pp. 321–328, 2002.

[90] D. Ghose, S. Panda, and P. Swain, "Prediction and optimization of runoff via anfis and ga," *Alexandria Engineering Journal*, vol. 52, no. 2, pp. 209–220, 2013.

[91] H. Hosoya and A. Hyvärinen, "Learning visual spatial pooling by strong pca dimension reduction," *Neural computation*, vol. 28, no. 7, pp. 1249–1264, 2016.

[92] N. Falco, J. A. Benediktsson, and L. Bruzzone, "Spectral and spatial classification of hyperspectral images based on ica and reduced morphological attribute profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 11, pp. 6223–6240, 2015.

[93] S. Wang and S. Liu, "Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm lda," *International journal of molecular sciences*, vol. 16, no. 12, pp. 30 343–30 361, 2015.

[94] F. Feng, W. Li, Q. Du, and Q. Ran, "Sparse graph embedding dimension reduction for hyperspectral image with a new spectral similarity metric," in *Geoscience and Remote Sensing Symposium (IGARSS), 2017 IEEE International*. IEEE, 2017, pp. 13–16.

List of Publications

1. Harsurinder Kaur, Husanbir Singh Pannu "*Blog response volume prediction using Adaptive Neuro Fuzzy Inference System*" 9th International Conference on Computing, Communication and Networking Technologies, IEEE [In Press]

2. Harsurinder Kaur, Husanbir Singh Pannu, Avleen Kaur Malhi "*A systematic review on challenges in machine learning: applications & solutions*" ACM Computing Surveys(CSUR) [Under Review]

3. Harsurinder Kaur, Husanbir Singh Pannu "*Anomaly detection survey for information security*" 10th International Conference on Security of Information and Networks, ACM, pages 251-258, 2017 [Published]