Data Mining Techniques for Social Network Analysis

Amina Siddique, Amna Kamran, Muhammad Aamir Gulzar, Rashid Noor, Ayesha Aziz and Dr. Kashif Munir
School of Computing
FAST National University of Computer and Emerging Sciences
Islamabad, Pakistan
Mail@nu.edu.pk

Data Mining Techniques for Social Network Analysis ***Abstract*—Social network analysis is very famous in these days. One of the main reason for this is bulk of data that is being generated by social networks. And when we extract data from any social network we have to use it for useful information. We preform some prepossessing on our data and then preform different machine learning techniques for the analysis of social network data to get useful information that maybe pattern generation, sentiment analysis, community detection, recommendation system etc. In this document we will apply different machine learning algorithms on labelled data-sets and then compare the results by calculating accuracy of different models.**

***Index Terms*—Data Mining, Social Networks, Social Network Analysis,Sentiment Analysis, Machine Learning**

I. Introduction

People across the globe are connecting together, sharing information on social networks and turning the world into a global village. Social networks connect people regardless of their nationality and age to share their opinion, express their feelings, emotions, hobbies, pictures, videos and sentiments using these online platforms[**introduction**]. Social network is a super set of events,ideas and functions where each node represents a unique entity. It has enabled stakeholders from all sectors to advertise,discover, analyse, gain knowledge, and enhance their companies using data from social media. Consequently, the role of social media for academia and industry is evident to research conducted in search of answers to critical questions.[**introd1**]

Social network analysis(SNA) is not a theory of sociology, rather it's a practical method of monitoring and tracking social networking sites(e.g Facebook, YouTube,Instagram,Twitter) are generating petabytes of data[**compressing**] milliseconds. SNA is used to extract and analyze valuable information with aims to detect communities[**intro2**], find following patterns [**pattern**],friend's recommendation[**recommad**] ,represent networks[**representNetwork**], and visualize social patterns[**socialpattern**]. The data from social networks is unstructured, precise, imprecise and uncertain data [**SOCIALMEDIA**]. Moreover, the social network produces huge amounts of continuous real time data. Statistical methods are incompatible to analyze this enormous data[**chen2012business**]. Therefore, the data mining techniques are conquering this issue.

Data mining is a process of extraction patterns, correlations and relationships in our data sets.Data mining techniques,methods and algorithms are used for analyzing and handling large volumes of data[**book**]. Like traditional Miners, data miners also extract useful information from a data set and convert it into structured data for further use. There are various data mining techniques which are used to extract data including Characterization, Classification, Regression, Association, Clustering, Change Detection, Deviation Detection, Link Analysis, Sequential Pattern Mining.To identify hidden patterns, supervised and unsupervised algorithms are used.

The process of determining whether a text or dialogue is positive, negative, or neutral is known as sentiment analysis. Many data mining algorithms rely on NLP techniques like syntactic parsing, part-of-speech tagging (POG), and other forms of linguistic analysis. Natural language processing tools, on the other hand, are often not effective in the social media domain. Social Network vocabulary is relatively brief and includes special words such as emojis, emphasis, and social media slang. Because traditional NLP cannot process such text, Deep Learning algorithms are used.[**sentiment**]

We select sentiment analysis for our experimentation and check which machine learning technique gives the best accuracy for sentiment analysis. Sentiment analysis is very famous nowadays because many companies wants to check their popularity among their customers on the basis of their feedback or reviews. We take a data set of an Amazon product reviews for the classification of customers sentiments either they like the product or not. For classification of sentiment we use five machine learning techniques and train our models on 80 percent training data and 20 percent we used for testing purpose. To check the accuracy of our models we use different accuracy measures like accuracy score, precision, recall and F1-score.

Some studies examined specific areas of data mining techniques used in social media. The purpose of this study is to investigate the available articles in terms of: a comparison of different data mining techniques, the strengths and weaknesses of the data mining techniques, the methodology used to extract social media data, the performance metric used for data mining techniques for social network analysis, and tools used to analyse different aspects and dimensions of data. We took

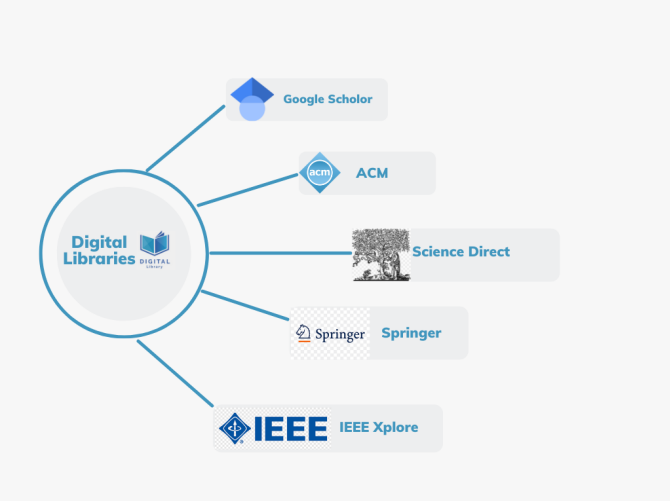

Fig. 1. Digital Libraries

sentiment analysis as case study and apply machine learning technique to check performance.

In this study, we conducted a survey using the Systematic Literature Review (SLR)[**r74systematic**] which consists of three stages: planning, conducting, and reporting. We began the planning phase by developing the research methodology, trying to identify the study sampling process, defining the quality measurement rules, documenting the data gathering strategy, and replicating the extracted data. The search strategy that we followed in this survey is explained in detail. We start by collecting research material from digital Libraries. The first search process included journals and Tier I social network related conferences from the aforementioned digital libraries, such as International Conference on International World Wide Web Conference (WWW),International Conference on Data Engineering (ICDE),ACM Conference on Online Social Networks (COSN),International Conference on Data Engineering (ICDE) and Advances in Social Networks Analysis and Mining (ASONAM). Furthermore to " " for key-phrases, we utilised Boolean operators (OR and AND) to restrict the search results. The following are the various search phrase used to extract relevant research papers.

- "machine learning" AND "data mining" AND "social media"
- "technique" AND "data mining" OR "technique" AND "social media"
- "fuzzy" AND "social media" AND "data mining"
- "machine learning" AND "data mining" AND "social media"
- "social media" OR "social network" OR"K-Means" OR "SVM" OR "support vector machines"
- "social media" OR "social network" OR "Apriori" OR "EM" OR "expectation maximisation" OR "PageRank" OR "AdaBoost" OR "KNN" OR "k-NN" OR "knearest" OR "k nearest neighbours"
- "technique" AND "data mining" OR "technique" AND "social media" OR "Naive Bayes," or "CART"

After the collection of some useful material we started reading of this material using three pass Approach. In the first pass we only go through the title, abstract,introduction, approaches techniques names and conclusion of the papers. In our second pass we studied the related work that have been done in the field of data mining for social network analysis and its different approaches. In the final pass we read all the implementation details of the papers that how a data mining technique affect the social network analysis. We studied different approaches of social network that are used to analyse using different techniques of data mining.

The purpose of this study is to investigate the available articles in terms of: a comparison of different data mining techniques, the strengths and weaknesses of the data mining techniques, the methodology used to extract social media data, the performance matrix used for data mining techniques for social network analysis, and tools used to analyse different aspects and dimensions of data.We took sentiment analysis as case study and apply machine learning technique to check performance.

II. Literature Review

In the area of social network analysis there is not a huge research done but now this area is emerging as we mentioned earlier due to high potential of data that is being generated from different networks. There are many machine learning techniques that can be used for specific type of data-sets to get better results. Some famous machine learning techniques used are Random Forest [**r9**],Naive Bayes [**r7**], K-Nearest Neighbor [**r6**],Decision Trees [**r8**], Logistic Regression [**r10**], Artificial Neural Networks (ANN) [**r4**] and Support Vector Machine (SVM) [**r5**]. In Table.1 we describe each of these techniques for some data-set with their accuracy for better understanding.

A. Performance Metrics

Performance metric shown in the Table.1 below is made on the basis of performance measures taken into account by different authors while contributing to the work in the field of data mining techniques for social network analysis. Different papers contain different sorts of metrics of evaluation of their work like some focused on classification, accuracy, precision, time etc and some were highly considerate about effectiveness, efficiency, and satisfaction. Survey papers compare work done by visualizing these metrics.

As the use of social networks are highly popular nowadays, a huge amount of data is generated by them, so the paper [**compressing**] is focusing on compressing the data and mining it. The major parameters for evaluation they are focusing on for this approach is effectiveness. The [**introduction**] discusses how accuracy and classification error rate is improved by the approach shared by them in this paper. The result also explains that Decision Tree (DT) gives the highest accuracy in different experiments than other ML approaches to figure out the depression.

TABLE I
PERFORMANCE METRICS

Paper	Effectiveness	Accuracy	Classification	Efficiency	Practicality	Precision	Time	Satisfaction
[**compressing**],[**multimodalApproach**]	yes	no	no	no	no	no	no	no
[**introduction**]	no	yes	yes	no	no	no	no	no
[**recommad**],[**representNetwork**],[**pattern**]	no	no	no	yes	yes	no	no	no
[**reviewbyML**]	no	yes	no	no	no	yes	yes	no
[**recommenderSystem**]	yes	yes	no	no	no	no	no	no
[**predictionn**]	no	no	no	yes	no	no	no	no
[**Enhanced**]	no	no	no	no	no	no	no	yes
[**r67Monetization**]	no	yes	no	no	no	no	yes	no
[**r68sentimentanalysis**]	no	yes	no	no	no	yes	yes	no
[**r69socialmediapost**]	no	yes	yes	no	no	no	no	no
[**r70fuzzylogic**, **r72twitterdatasentimentanalysis**]	no	no	yes	no	no	no	no	yes
[**r71deeplearning**]	no	yes	no	no	no	yes	no	yes
[**r73**]	no	yes	yes	no	no	no	no	no

In the present time of big data, collection and generation of very large amounts of valuable data is comparatively easy. The paper [**recommad**] focuses on recommending friends so the parameters they focused on are efficiency and practicality. Other papers are also displayed in the similar way in the table with the corresponding performance evaluation results discussed.

In this paper we discussed sentiment analysis in detail ,which is one of the important domain of social network analysis. For the performance evaluation of our experiments we used different accuracy measures like accuracy score, recall, precision and F1-score. We selected five machine learning techniques for the sentiment analysis of customer's tweets on amazon product. We use accuracy, precision, recall and F1-score as performance measures for these five techniques on customer reviews dataset is presented in Table 3.1.

Some of the performance metrics analysis using the result of different journals is explained in the following table.

TABLE II
COMPARISON TABLE OF DATA MINING TECHNIQUES

Paper	Results
[**r37sentimentanalysis**]	The analysis show that the model performs well and is effective. When the willingness level was 0–30 percent, the average unwanted rate was 43.32 percent with a standard error of 1.79 percent, according to the paper's example. For a willingness level of 71 percent–100 percent, the mean undesired rate was 4.84 percent. They discovered that the higher the level of willingness, the lower the unwanted rate, and vice versa.
[**r39sentimentanalysis**]	In comparison to a flat classification, the hierarchical classification technique improves performance significantly. The classification results are considerably improved when semantic orientation (SO) characteristics are combined with Bag of Words (BoW) features. The accuracy of 55.24 percent is achieved.
[**r38sentimentanalysis**]	The evaluation results of their model match well with practical situations.
[**r75crosslingual**]	Experiments on 18 cross-lingual sentiment categorization tasks show that the proposed approach is more effective and powerful than earlier studies.
[**r76thai**]	Higher accuracy than SVM and Nave Bayes, but lower than Maximum Entropy. Original sentences have higher accuracy than shuffled sentences.

B. Research Methodology

Online Social Networks are famous nowadays due to huge amount of data they are producing. We need to get this online data and extract useful information from this data. There are many approaches to extract and process data. Before discussing that it is important to discuss the methodology which is mostly used for this purpose. There are four basic steps; In first step we have to select any social network platform for data gathering. In second step we have to collect and extract useful data from previous selected platform using different techniques that we will discuss later on. For quick view of these techniques you can see Table.2. In third step we will use different algorithms to evaluate social networks and data. In last step we will analyze the results and on the basis of these results we can do future analysis. All the steps are mentioned in Figure.1 below.

If we have a data first of all we have to check that this is related or unrelated data. Related data always have some relation or linkages with each other and this type of data is called Graph structured data. [**categorize**] In other case if data is unrelated this is called Unstructured data.In related data we can identify relation for this we can do social interaction analysis and social data analysis. [**categorize**] But for the case of unstructured data first we have to do structural analysis of the data. For your better understanding we have given below the Hierarchical tree for different types of social data and their analysis. trees [level 1/.style=sibling distance=7 cm,level 2/.style=sibling distance=4.5cm] Data[edge from parent fork down] child node Graph Structured (Linkage Data) child nodeSocial Network Analysis child nodeSocial Interaction Analysis child nodeCascading child nodeInfluence child nodeNetwork Influence or Behaviour child nodeSelf Organization child nodeSocial Data Analysis child nodeSentiment Analysis child nodePrediction child nodeTrending Topic Detection child

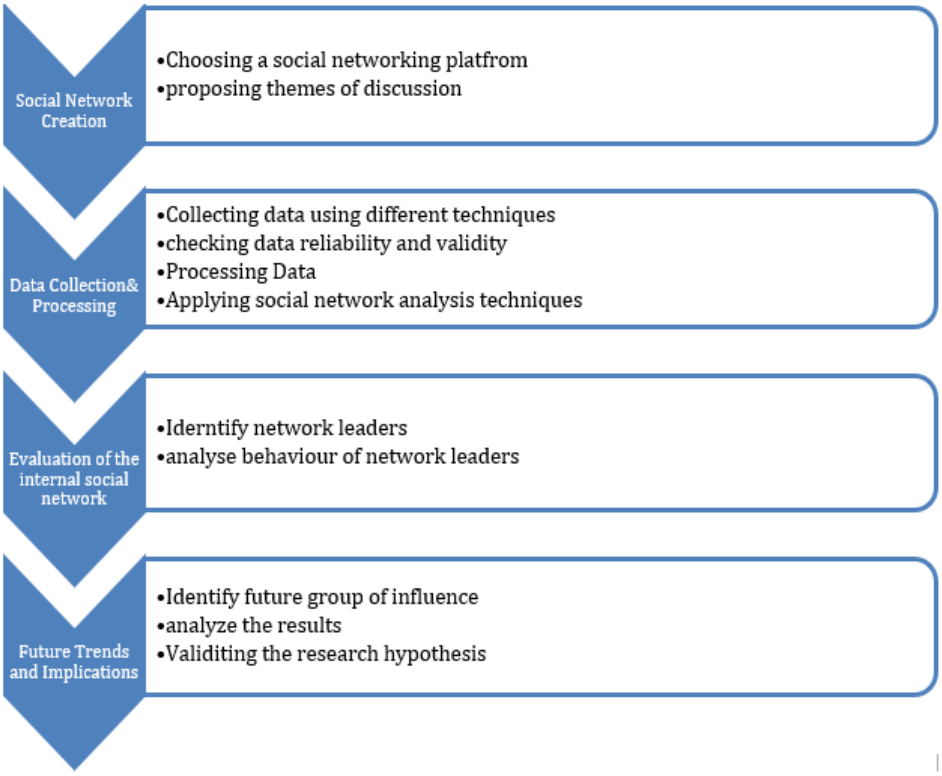

Fig. 2. Proposed Methodology

nodeLocation and Real Events Detection child nodeSocial Recommendation Systems child node Unstructured (Content Data) child nodeBig Data Analytic child nodeStructural Analysis child nodeTopology Characterization child nodeUse/Functionality Characterization child nodeAnomaly and Fraud Detection child nodeRepresentation Models ;

After classification of data now we can easily classify data extracted from our social networks. Now we have to discuss approaches that we can use for analysis of data and its results evaluation. We have discussed different techniques for the analysis of given data according to their specific approaches in Table.2 below.

In above Table 2.2 we discussed different approaches of social network analysis and the machine learning techniques that are mostly for the implementation of these approaches. For better understanding of different approaches and techniques we make comparison table for some social network approaches and the techniques that we mostly use for them with their strengths and weaknesses, which can be seen in Table 2.5.

Moreover, in this paper we implemented the sentiment analysis approach using these five machine learning techniques Logistic Regression, Support Vector Machine, Artificial Neural Network(MLP classifier),Decision Tree and Random Forest. We perform sentiment analysis on an Amazon product reviews dataset that is available on Kaggle website. This dataset contains customer tweets data for the product which they buy from amazon website. We perform classification to check either the review is in favor of the product or not. We divided our dataset into 80:20 ratio eighty percent for training and twenty percent data for testing. After the training of our model we predict some tweets and to check its accuracy by

TABLE III
RESEARCH METHODOLOGY

Reference	Approach	Technique
[**r21influence**]	Influence Propagation	Centrality measure, Graph Theoretic
[**r22communitydetection**] [**r23communitydetection**]	Community detection	Hierarchical and vertex clustering
[**r24expertfinding**, **r25expertfinding**, **r26expertfinding**]	Finding Expert Finding	Probabilistic Latent Semantic Analysis (PLSA)
[**r27linkprediction**, **r28linkprediction**, **r29linkprediction**]	Link Prediction	Markov-chain model, SVM, K-Nearest Neighbors
[**r30recommendersystem**, **r31recommendersystem**, **r32recommendersystem**]	Recommender system	CF (Collaborative filtering), Clustering
[**r33trustPrediction**, **r34trustPrediction**, **r35trustPrediction**, **r36trustPrediction**]	Trust Prediction	EigenTrust, VoyeurServer, Matrix factorisation
[**r37sentimentanalysis**, **r38sentimentanalysis**, **r39sentimentanalysis**, **r40sentimentanalysis**]	Sentiment Analysis	Socioscope, Hierarchical classification technique
[**r41opinionanalysis**, **r42opinionanalysis**, **r43opinionanalysis**]	Opinion Analysis	Aspect Based, Support vector regression

comparing the predicted results with the actual results from our testing data. We use different accuracy measures to check the quality decision of our overall predictions like accuracy score, recall score, precision and F1-score.

C. Comparison Of Approaches

There are many approaches in social notwork analysis which we can use for classification, sentiment analysis, friend suggestions and for many other objectives. We discussed some of these approaches in Table.3 below with their strengths and weaknesses.

D. Evaluation Tools

In social network analysis, there are multiple tools used to analyze different aspects and dimensions of data, by applying different algorithms to extract useful and necessary details. Below is the comparison about the different software tools which are being used widely to for different purposes.[**tools1**]

Different tools have different limitations and as per the requirements of the data, constraints, volume, nature of the information, and presentation of the data. [**tools4**] Above mentioned software applications are for analysis and visualization of networks, these are to provide the user powerful visualization tools, efficient algorithms for analyzing networks and abstraction by factorization of network into smaller networks.[**tools4**]

Furthermore, there are several software packages especially designed for investigating other structures than social networks (e.g. network text analysis, ecological network analysis, simulation of evolutionary environments etc.).[**tools4**]

Gephi is an independent co-relational software which study the behaviours of nodes and edges of any network structure by visualizing the patterns, it does support classic data mining. Gephi allows to study the connectedness of any network.[**tools1**][**tools2**][**tools3**][**tools4**][**r20educationaldatamining**][**gephi-cite2**] Pajek is general is a tool to analyze the graph network with certain patterns of nodes and edges and mutual relationship. To visualize the large network in the form of graph. There are different range of metric like portioning schemes, cliques, clusters, network components etc.[**tools1**][**tools2**][**tools3**][**tools4**][**r20educationaldatamining**][**pajek-cite1**] Node[XL] is a free tool, for analysis and visualization of network data using media and graphs. Computation in Node[XL] is quite slow as compare to Gephi and Pajek, so many research efforts has been applied to increase the efficiency of the this tool.[**tools1**][**tools2**][**tools3**][**tools4**][**r20educationaldatamining**][**networkx-cite3**]

NetworkKit is a python tool, which is high performance tool to sketch the high level graphs, parallel execution is allowed in this software tool which is why this is quite efficient and significant tool for research purposes in social network analysis, mainly used for heavy data requirements. Python libraries in this tools do provide the interactivity to some extent for better understanding of the data.[**tools1**][**tools2**][**tools3**][**tools4**][**r20educationaldatamining**][**networkx-cite3**] Tulip is a information visualization framework which is being used for the representation and depiction of relational data of larger network. Vision is a tool in java language for for research purposes in network analysis, it contains interactive graphical user interface, available in window, Linux. More importantly import, export of network data is available on this software tool.

Above table is the illustration for metrics which are being measured and analyzed using different social network analysis software tools.[**tools1**] Few performance metrics are being discussed in the table to compare the performance of different metrics in different software tools in social network analysis. Which are being analyzed and described in different papers of different authors.[**tools2**] Degree, closeness, density, page rank, and network diameter are few performance metrics which are being analyzed using different data-sets in different experiments. Different social network analysis tools do have different capabilities to analyzed and illustrate different performance metrics.[**tools1**][**tools4**]

III. Case Study: Sentiment Analysis

A. Description

As population of the world increases social network are also growing with the passage of time [**r91sentiment**]. People wants to connect with each other to share information with each other. There is huge production of data on daily basis, which can be used to for analysis of people interest, sentiments, recommendation and behaviour [**r92sentiment2**]. To extract this useful information from raw data that is collected from any social network machine learning is very useful. If there is some noise in the data first we have to preprocess this kind of data before analysis. So there are many techniques that can used for data preprocessing, which are not included in scope of our paper so we will not cover details of data extraction and preprocessing techniques.

After cleanliness of data we have to perform analysis on it and here machine learning comes into play. There are many techniques that can be used for analysis of social network but it also depends on our data-set [**r93sentiment3**]. We have to choose that machine learning technique for our data-set, which should be able to get better results. Here in our experimentation we use dataset "Sentiment analysis on Amazon Reviews". This dataset is available on Kaggle website. In this dataset we have Amazon product reviews text and we have to analyse either tweet is in favor of product or not. Machine learning models that we use for our experiment are Logistic Regression, Support Vector Machine, Artificial Neural Network(MLP classifier),Decision Tree and Random Forest. To check either this technique is good for our data-set or not we calculate accuracy of the model.To measure model accuracy we use accuracy, precision, recall and f1-score.On the basis of accuracy, time and complexity of the different machine learning techniques we can decide best technique for our data-set.

IV. Methodology

In our methodology portion we use the same dataset "Sentiment analysis on Amazon Reviews" and choose five machine

TABLE IV
COMPARISON TABLE OF DATA MINING TECHNIQUES

Technique	Strength	Reference	Weakness	Reference
Graph Theoretic	Identify communities and users with most influence	[**r44graphlimitations**]	Overemphasis, Computational limitations	[**r44graphlimitations**]
Hierarchical clustering	It can be used for large no of clusters	[**r45HC**]	Overestimation	[**r45HC**]
Probabilistic Latent Semantic Analysis (PLSA)	Useful for sentiment classification, Reduce complexity, High feasibility	[**r46PLSA**, **r47PLSA**]	Limited to Topic classification	[**r46PLSA**]
SVM	• Best for multidimensional network and offline clustering • Most effective methods for resolving classification problems	[**r45HC**, **R62mining**]	Link prediction is challenging task	[**R63mining**]
K-Nearest Neighbours	• High classification Accuracy, Time Efficiency, Can be used on Noisy Data. • Relatively simple and most exclusionary pattern matching classifiers.	[**r49KNN**, **R57decision**]	Not efficient for large datasets, data with high dimension and missing values	[**r49KNN**, **r58nearest**]
Collaborative Filtering (CF)	Can be used for dynamic classification	[**r50CF**]	Not efficient for noisy data	[**r50CF**]
Matrix Factorisation	High accuracy for recommendations like friend or product suggestions	[**r51MatrixFactorization**]	Not suitable for heterogeneous data	[**r51MatrixFactorization**]
Support vector regression (SVR)	Useful for prediction of user interactions	[**R52SVR**]	complex and Time consuming	[**R52SVR**]
ANN	Increased capabilities that aided high-dimensional data processing.	[**r61ann**]	Stimulate maps of lower grade than the kernel version's maps.	[**r61ann**]
K-Mean	• Appropriate for a fixed variety of groups with unfamiliar features based on various user-defined factor. • Perform well as it comes to identifying a small clusters .	[**r55clustering**, **r56classifying**]	Quality of found clusters rapidly weakens as clusters rises.	[**r55clustering**]
Fuzzy	Modeling with ambiguous ways of social reasoning is a specialty, and it takes into consideration the unpredictable aspect of human cognition.	[**r59social**]	Manual process of handling the semantic fuzzy rule through an offline approach necessitates knowledge of semantic web and fuzzy systems.	[**r59social**]
XMeans	The graph API is used to collect data. It produces effective results on sorted data.	[**r60clustering**]	When the user understands how many clusters are required, only than this algorithm is useful. It doesn't work well on hierarchical clustering.	[**r60clustering**]
Association	It's good to use for the patterns of same variables, also known as relation technique, good to show probability.	[**data&mining**, **data˙mining**]	Too many variables and for non experts it a heck to explore and to understand the real findings..	[**data&mining**]
Decision tree	It is good to use as a selection criteria, for data which is selection specific. Simple two state solution as a decision tree.	[**data&mining**, **data˙mining**]	Unstable behaviour and nature of this technique. Less reliable when deciding on continuous changed variables.	[**data&mining**]

TABLE V
COMPARISON TABLE OF SOFTWARE TOOLS

Program	Package	Parallel	Execution	Graphics	Capabilities
Gephi[tools1, tools2, tools3, tools4, r20educationaldatamining][gephi-cite2]	Java	No	Low	Medium	Business
Pajek[tools1][tools2][tools3][tools4][r20educationaldatamining][pajek-cite1]	Java	No	High	Low	Business/ Academic
NodeXL[tools1][tools2][tools3][tools4][networkx-cite3]	C#.net	No	Medium	Medium	Business
NetworKit[tools1][tools2][tools3][tools4][r20educationaldatamining]	Python	Yes	High	High	Business/ Academic
Tulip[tools1][tools2][tools3][tools4][r20educationaldatamining]	C++	No	Low	Low	Academic
Vison[tools1][tools2][tools3][tools4][r20educationaldatamining]	Java	No	Medium	Medium	Academic

TABLE VI
COMPARISON TABLE OF SOFTWARE TOOLS

Program	Gephi	Pajek	NodeXL	NetworKit	Tulip	Vision
Degree[tools1][tools2][tools3][tools4][r20educationaldatamining]	Yes	Yes	Yes	Yes	No	Yes
Betweenness[tools1][tools2][tools3][tools4][r20educationaldatamining]	Yes	Yes	Yes	Yes	No	No
Closeness[tools1][tools2][tools3][tools4][r20educationaldatamining]	Yes	Yes	Yes	Yes	Yes	Yes
Density[tools1][tools2][tools3][tools4]	Yes	Yes	Yes	Yes	Yes	No
Page Rank[tools1][tools2][tools3][tools4][r20educationaldatamining]	No	Yes	Yes	Yes	No	Yes
Network Diameter[tools1][tools2][tools3][tools4][r20educationaldatamining]	Yes	Yes	Yes	Yes	Yes	Yes

learning models for training and testing of data to check which model performs best. Five models that we choose are Logistic Regression, Support Vector Machine, Artificial Neural Network(MLP classifier),Decision Tree and Random Forest. We downloaded the train and test files of data separately from Kaggle website then we write a python function to load the text and labels in our program. After this we do some prepossessing on our data firstly we do data cleaning by removing punctuation's and stop-words from the text. To convert the text into vectors we use count-vectorizer tool. After all the prepossessing of data and converting it into vector we split our data 75 percent data for training of model and 25 percent for testing.

Decision Tree Decision tree classifier is a supervised classification algorithm which selects the best features from the data on the basis of information gain and then takes decision on the basis of selected features [**r94DecisionTree**, **r95DecisionTree2**, **r96DecisionTree3**]. After the classification of data on one best feature it calculates the entropy and check the impurity of the leaf nodes. If any leaf node has zero impurity this is declared as a final leaf node otherwise in case of impurity we decide to perform further classification on the basis of some other features. We use decision tree classifier with its default parameters, which have default 'gini' criterion (this is function that is used to check the quality if split), default 'best' splitter, no maximum depth, 2 minimum samples split, 1 minimum samples leaf, 0 minimum weight and random state . With all these default parameters we achieve almost 63 percent accuracy on this dataset. We changed some of its default values but this was the best accuracy opn the default values. We perform post pruning on the tree and retrain the model with 0.012 cc-alpha value and then test the model. This time the accuracy of model increase one percent in this the highest accuracy we achieve from decision tree classifier is 64 percent.

Logistic Regression Logistic Regression is a supervised machine learning algorithm. We can even use this for multi-class using one-vs-rest rule [**r97Logit**, **r97Logit2**]. We implement this algorithm on our reviews dataset after all the preprocessing and vectorization techniques with its default 'l2' penalty and different C (C is a regularization parameter) values. We achieve around 73 percent accuracy for 0.5 value of c.

Multi-layer Perceptron classifier (MLP) This is an artificial neural network algorithm [**r102MLP**, **r105MLP4**], which optimizes the log-loss function using LBFGS or stochastic gradient descent. This has inputs layers, hidden layers (Neurons), and output layer [**r103MLP2**, **r104MLP3**, **r106MLP5**]. Inputs are given in the dataset and the output layers also depends on the number of classes that we want in our case there will be only two output layers. The hidden layers are also called neurons there are very important for model training and it depends on us how many no of hidden layers we want it also depends on the datset. More number of hidden layer mostly increase accuracy of the model but it is very complex to solve and requires more computation. We use MLP model with 1 random state, 300 number of maximum iterations and default remaining parameters like 'Relu' activation function, 'adam' solver, 0.0001 alpha and constant learning rate. The maximum 75 percent accuracy achieved from MLP classifier.

Support Vector Classifier (SVC) This is linear model of support vector machine which is used for classification [**r107SVC**, **r108SVC2**] and this can also be extended for multi-class classification by using one-vs-reset scheme. This model a little differently from others because this is not trained using all the samples of training data this only takes two points as support vectors, which helps for the classification of all other points. SVC always tried to maximize the distance between these two support vectors because this is helpful for easy classification [**r109SVC3**]. If any points lies in the

centre of these support vectors then we have to calculate its distance from the origin and for this purpose we use the function soft margin classifier for the final decision about that particular point [**r110SVC4**]. We use this algorithm with the 'auto' gamma value while all other parameters are default. We achieved 44 percent accuracy by using SVC classifier.

Random Forest Random forest is similar to decision tree algorithm but it makes more than one trees by taking the subsets of the features and samples of the data [**r98RF**, **r99RF2**]. After this it takes final decision on the basis of maximum votes in case of classification problem but in case of regression it takes sum of the averages of all the trees and then take decision on the basis of value [**r100RF3**, **r101RF4**]. We use random forest classifier with 20 maximum depth and 0 random state etc. All the other parameters like criterion,min-samples-split, class weight etc are set to be default. We are able to achieve a maximum of 68 percent accuracy using this algorithm.

We trained and test all the above mentioned models with the same dataset to compare their performance. For this purpose we uses these four performance metrics accuracy, precision, recall and F1-score.

Accuracy basically identifies the number of correctly identified data samples over the total number of samples. The formulae to calculate the accuracy is given below,

Accuracy = TP + TN / TP + FP + TN + FN

- **TP(True Positive)** These are those data samples which are predicted as a positive and their actual label is also positive.
- **TN (True Negative)** These are those data samples which are predicted as a negative and their actual label is also negative.
- **FP(False Positive)** These are those data samples which are predicted as a positive and their actual label is false.
- **FN (False Negative)** These are those data samples which are predicted as a negative and their actual label is positive.

Precision basically identifies the number of positive correctly identified data samples over the sum of positive correctly identified and positive wrong identified data samples. Precision should be ideally 1 high for a good classifier. The formulae to calculate the precision is given below,

Precision = TP / TP + FP

Recall basically identifies the number of positive correctly identified data samples over the sum of positive correctly identified and negative wrong identified data samples. The formulae to calculate the recall is given below,

Recall = TP / TP + FN

F1-Score is basically a metric which takes into account both the precision and recall. The formulae to calculate the F1 Score is given below,

F1-Score = 2 (Precision * Recall) / (Precision + Recall)

After calculating all these performance metrics we compare these scores of all the five machine learning models which can be seen in Table 3.1.

A. Results

Social network analysis is a very wide area of study if we want to compare machine learning model performance for social network we first have to specify our area what we wants to predict like in our experimentation we do sentiment analysis on social network data using different machine learning models. To measure the performance of models and their comparison we calculate accuracy, precision, recall and f1-score. The results of all these accuracy measures are mentioned in Table-II.

TABLE VII
PERFORMANCE METRICS FOR SENTIMENT ANALYSIS

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.744	0.69565	0.8	0.74418
Decision Tree	0.632	0.5675	0.75	0.64615
ANN (MLP)	0.752	0.6923	0.8035	0.7438
Support Vector Machine	0.448	0.448	1.0	0.6187
Random Forest	0.6896	0.71165	0.66276	0.68633

As of our results if we compare Logistic Regression, Support Vector Machine, Artificial Neural Network(MLP classifier),Decision Tree and Random Forest for sentiment analysis Artificial Neural Network (MLP classifier) performs better than other with highest accuracy value and F1-score. On the second number logistic regression in this case also performs almost equally to the ANN with second highest accuracy and other measures are also good. For exact values you can see Table-II. We can conclude that in social network analysis for sentiment analysis Artificial Neural Network (MLP classifier) and Logistic regression are the best machine learning models.

V. OBSERVATIONS AND FINDINGS

Most of the data mining techniques in social network analysis (SNA) are using classification techniques to generate the desired results with maximum ratio of accuracy. Number of classifiers are developed over the period of time to speed up the process, as it is such a laborious work to generate the results of huge data-sets in limited time. Over the last years social network analysis shown significant impact on different fields of life. To analyze the social media trends and sentiment different techniques are being used such as clustering, classification, etc. different techniques have their different benefits and limitations, as per the desired results different techniques are being used in different scenarios.

According to our observations and literature review there is no comparison between the different techniques that can be used for a specific type of approach in social network analysis. Mostly the accuracy of machine learning models don't depends upon the approaches it depends on the dataset given, which we use for then training of our models. We were unable to classify some techniques that are specially used for some specific type of approach rather than the Graph Neural Networks that are mostly used for community prediction and gives efficient results as compare to other models.

Sentiment analysis is very popular because most of the companies are concerned about their popularity among the people and they want to predict the future of their products. So there are many techniques discussed and suggested in many papers for sentiment analysis but most of them are on textual data. There are many audios and video channels where customer give their opinions related to any product or brand so how we ca n perform sentiment analysis on the live stream data this work is very less. Another thing that we observe is that mostly research work is related to sentiment analysis on the textual data like data of tweets, any product reviews, social network posts and articles etc but there is not enough work on the sentiment analysis of live video streams and audio taps, which are very popular nowadays. Mostly people record video or share their opinion related to any product or person in the short videos so there should be some research in this area for the future prediction of companies according to its customer opinions.

Sentiment analysis can further be extended to extract the hot topics from the discussion of the customers related to any product or service. For example if a customer buys a product from amazon he/she will leave o review and the review maybe a positive or negative according to customer experience. In case the customer leave a negative review related to any product or service he/she will write some reasons why the product didn't meet the expectations of customer. If we develop our sentiment analysis model strong that it will classify customer review either positive or negative and if its negative it should extract the reasons or keywords that customer mentioned in the review. This data can be further used for the improvement of the product by the company and this can reduce negative reviews in the future. Moreover, we can extract the keywords from positive reviews and this can be used for marketing purpose of company products.

A very straightforward and direct unsupervised classification technique can be used to identify the objects, adjectives, and other part of the speech in a sentences of a given text. Then using the semantic orientation we can approximate the phrase then we can classify using the average semantic orientation. Keywords can play the important role in this experimentation of the phrases. Semi-supervised learning is an activity with targeted goals and targets, but unlike unsupervised learning it can evaluated very specifically. It would be working in a way that we can train a model with specific classifier with data of negative and positive sentiments. In semi-supervised we use polarity detection as as label propagation, it works well on graph base models. While implementing the clustering techniques we can use the supervised learning classification where basis of data are well established but patterns are unknown. We can also implement this where we have already identified the data organization. Supervised classification use the combination of semantics orientation with diverse similarity or dissimilarity of facts of label couple of adjectives. On the basis of sentiment classification we can access our findings and results about the nature of the results and metrics which are under observation. It is critical to understand the nature of the data and then implement the right set of tools and techniques.

In this survey paper, we observed when different techniques can be used in social network analysis. Different data mining techniques, on the basis of their strengths and weaknesses, are applied to different data sets. We have observed that major data sets available for social network analysis are labeled, so supervised machine learning techniques are used. Majorly, logistic regression, random forest, and artificial neural networks are used in supervised ML techniques, and DB Scan and K means plus plus are used in unsupervised ML techniques.

VI. Conclusion

From our studies and research, we concluded that Social Network Analysis (SNA) lies between different For example, in computer science, mathematics, etc. This diversity led to the potential for analysis in In the field of SN, which paved the way to the inception of new methodologies and techniques over so many tools. This vast opportunity in SNA resulted in the development and application of new and diverse tools. in industry and academia to understand the complexities of networks using graphical representation. and visualisation of data. Different software is being used to perform tasks on tonnes of data with millions Nodes can perform various operations related to graph metrics, such as degree, centrality, and network size. diameter, etc. Tools are being used as per the need and requirement of the algorithm which ought to be performed with the help of different libraries for the computation and analysis of the data. It is concluded. According to our research, there is a wide range of options available in the field of data mining in SNA today. and a set of tools with rich accuracy of results and functionalities to choose as per our need for data and The desired outcomes are being used in both industry and educational research. purposes. Finally, at the moment, the main challenge that we are facing is exploration. of graphs with high-level visualisation with different orientations.

Our conclusions is summarized in following points:

- Data mining strategies have diverse strengths and drawbacks, therefore the choice of approach is depending on the type of informative data required.
- Supervised machine learning techniques are majorly used for social network analysis as the data of SNA is mostly labeled and unsupervised datasets are limited.
- In sentiment analysis, Artificial Neural Network(ANN) and Logistic Regression (LR) performs better with high accuracy value and F1 score.
- Logistic regression, Random forest and Artificial Neural Network models are better in performance in supervised machine learning.
- DBScan and K Means plus plus algorithm performs better in unsupervised machine learning.

Our research has concluded that today, just in the field of data mining in SNA, we have a vast diversity of options and sets of tools with rich accuracy in results and functionalities to choose from as per our needs for data and algorithms to

get the desired results, which are used both in industry and for educational research purposes. At the moment, the main challenge that is being encountered by us is the exploration of graphs with high-level visualisation with different orientations.

VII. Future Work

Future research can be conducted on the other areas of social networks like community detection, behaviour detection and influence detection to check which machine learning models perform best for these kind of researches in social networks. Some other machine learning models can also be considered for better analysis as in our experimentation we only used five supervised machine learning algorithms. Mostly, the work done in sentiment analysis is on textual data there maybe further studies conducted for sentiment analysis on the videos and audio data. As in social network analysis live data matters more than the old data because data is dynamic changes after each interval of time. So the model that is trained in online learning performs better than the model which is trained on batch learning because this is more useful for future predictions.

M Aamir Gulzar received the B.S. degree in computer science from PMAS Arid Agriculture University, Rawalpindi, Pakistan, in 2021. He is currently pursuing M.S. degree with the School of Computing, FAST National University of Computer and Emerging sciences, Islamabad, Pakistan. His research interests include social networks, artificial neural networks and other machine learning techniques.

Rashid Noor received the B.S. degree in software engineering from Riphah International University, Islamabad, Pakistan, in 2021. He is currently pursuing M.S. degree with the School of Computing, FAST National University of Computer and Emerging sciences, Islamabad, Pakistan. His research interests include graph neural networks, social network analysis and data modelling with advance algorithms, systematic analysis of software development techniques.