IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306225|

Data Mining Techniques in Social Media

Dishank Gawas, Prasad Chavan

Department of MCA, Vivekanand Education Society's Institute of Technology(VESIT),Chembur, Mumbai, India

Department of MCA, Vivekanand Education Society's Institute of Technology(VESIT),Chembur, Mumbai, India

ABSTRACT: Social Media which is one of the most major online platform in this modern era which connects billions of user from all over the world and contains a huge amount of heterogeneous data .The process of extracting meaningful pattern from heterogeneous data is known as Data Mining or Social Media mining. Social Media mining involves extracting, analyzing, and representing useful patterns from data derived from social interactions. This paper explores various data mining techniques applied to social media, focusing on their roles, challenges, and applications,evaluates their effectiveness and determines the most optimize and effective method. .We refer to multiple references and research papers to present a comprehensive analysis.

KEYWORDS: Data Mining Techniques, Data Mining in Social Media,Social Media Mining

I. INTRODUCTION

Social Media which is one of the most major online platform in this modern era which connects billions of user from all over the world.There are many various online platforms like Facebook,Instagram,Twitter and LinkedIn which create a large amount of data in every fraction of time.The data can be in any format text,video,audio,images and gif . The amount of heterogeneous and vast amount of data which is created is an opportunity and also difficult as well as challenging part for data analysis.

To extract meaningful and insight pattern from the data various data mining techniques are used. Data Mining is a process where meaningful patterns are discovered and analyzed from a large amount of datasets. The increasing growth of social media in 20th Century has raised the importance of data mining techniques.These techniques not only help in analyzing user behaviours and preferences but also in analyzing and predicting trends,detecting unethical activities and increasing user services.These techniques helps organizations,businesses and researchers to gain valuable and understandable patterns that can help one and improve their ability to make strategic decisions and innovations.

Objectives

1. To evaluate different data mining techniques used in social media to analyze user behavior and data.
2. To determine the optimal,best data mining techniques and least optimal data mining technique.

II. DATA MINING TECHNIQUES IN SOCIAL MEDIA

In this section, we have deep dived in to data mining techniques used in social media.So through a diagram (Fig1). we have classify the different data mining techniques and we have also explained it separately below.

List Of Data Mining Techniques are:

1. Naive Bayes
2. Support Vector Machine
3. Decision Tree
4. K-Means Clustering
5. Linear Discriminant Analysis
6. Maximum Entropy
7. Apriori Algorithm
8. Neuro Fuzzy Algorithm
9. Hidden Markov Model

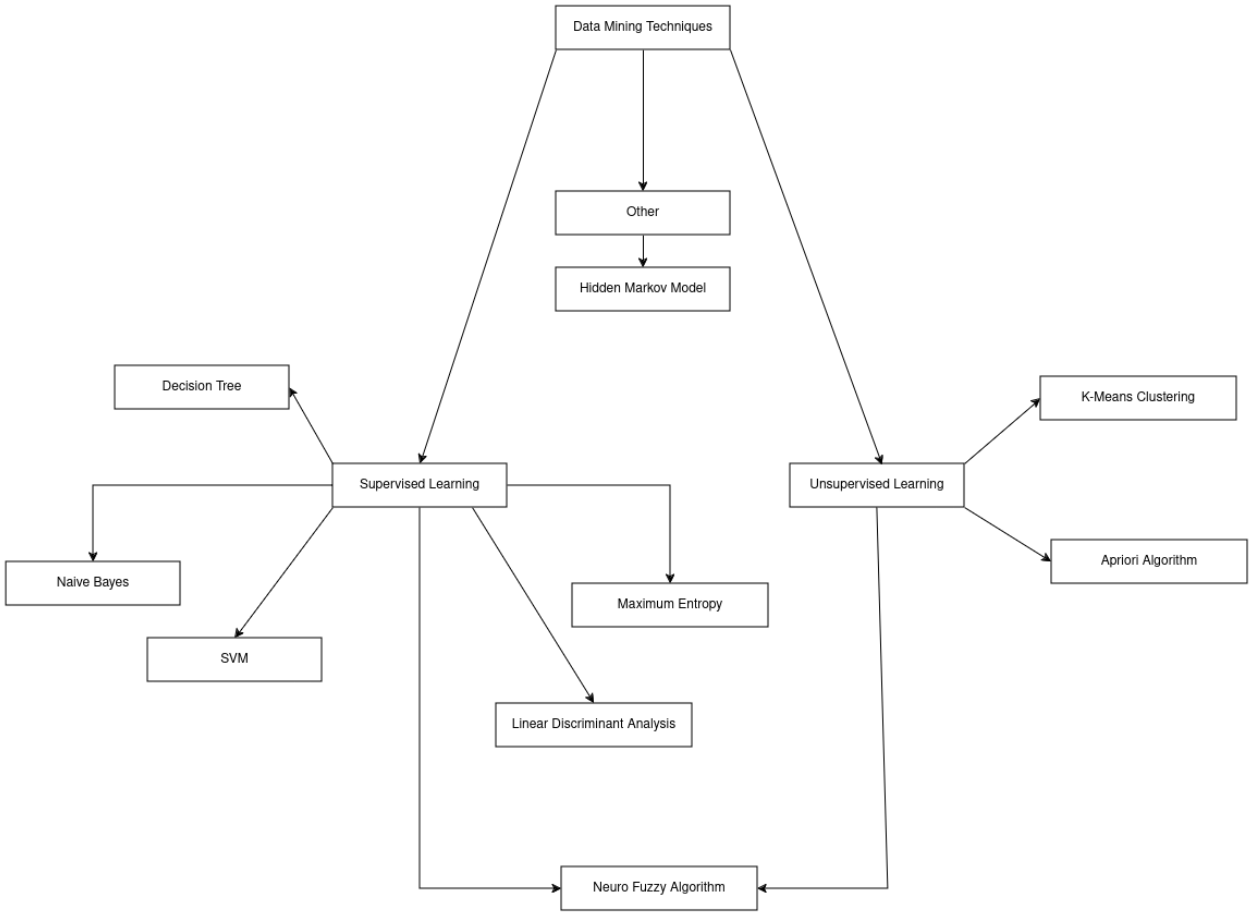
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306225|



(Fig 1)

Naive Bayes Classifier is one of the classification techniques in data mining.According to our researches we found out that naïve bayes text classifier was widely used to bifurcate between various type of content on social media. For Eg: Whether a particular discussion or content are closely linked to real world events or not,Naive Bayes was mostly used for real time analysis rather than predicting future predections.While explicitly it's not mentioned that it predict future events but it helps in identifying topics,trends and content in real time.While researching we found out that a naive bayes text classifier was trained to distinguish rapid-growing real world events and non events content on twitter regarding politics,elections or any other real world events occurring.Naive Bayes Classifier is also mostly used in sentiment analysis to analyze user's opinions on social media.

Support Vector Machine falls under the category of supervised machine learning algorithm.It is majorly used for classification and regression. Support Vector Machine (SVM) has demonstrated efficiency in training on Twitter hashtags metadata to predict the political inclination of Twitter users. The reason why SVM is more optimal solution because it finds an line which is optimal(hyperplane)which can classify the data.SVM is best option.Social Media is flooded with tons of information and hastags trending and since it becomes difficult to classify the unstructured data.SVM model works fast and can handle classification without slowing down.It was studied that SVM has an accuracy of about 95-98% as compared to Naïve Bayes Classifier.Whereas Naive Bayes classifier has and accuracy of about 75-80%.SVM is also used for fake news detections in social media.

Decision Tree Classifier is a supervised learning algorithm and classification technique.It is mostly used to solve classification problems.Decision Tree classifier follows a tree like structured classification where the internal node represents the characteristics of dataset, branches represent the rules for making decision and every leaf represent the

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306225|

desired output.Decision Tree Classifier uses CART algorithm.To know how social media was perceived by users took a survey and collected data from students of university who actively uses Facebook.They collected their demographic information and the amount of time they are active on facebook and also they asked students to rate facebook usefulness for educational purpose.After survey they found out that almost 15-20 students from 550 didn't provided enough information so they removed their information from data.After analyzing researchers discover underlying and meaningful pattern on how students use facebook for educational purpose.Although Decision Tree Classifier is not widely used for social media mining but it can be taken in consideration for classification.

K-Means Clustering is an unsupervised learning algorithm and falls in clustering technique.The objective of K-Means is to group data points which are similar and discover underlying patterns and trends.A target number K is defined which indicates the number of centroids included in dataset and also it represents the center of cluster.The K-Means Clustering identify k centroids and set every points to the nearest cluster keeping centroid as minimum as possible.K-Means clustering is not that much used widely but it can be used in recommendation system and categorizing contents.In social media K-Means can classify different trends and posts in categories.For Eg If a user searches for specific thing like happy post or a book then K-Means Clustering will classify those content in cluster through text analysis and suggest them similar content on user-based preference and surfing behaviour

Linear Discriminant Analysis (LDA) is a dimension reduction technique.It is mostly used in ML to classify multiple classes problems.When the primary goal is to classify multiple classes who have N numbers of features and characteristics Linear Discriminant Analysis is the best suited data mining technique for this purpose.As we cant classify two or more classes base on a single common feature or characteristics which will ultimately result in overlapping.So to overcome this issue number of features are increased and then through LDA the 2-D plane is reduced to 1-D due to which it becomes lot more easier to draw a straight line and separate two classes.LDA is used in social media for user identification and predictions.In social media, LDA can easily predict and can include those features who are most likely to engaged with same or specific content on social media platform by group of users.LDA method allows identify user behaviour and it becomes easy to target them with relevant contents.

Maximun Entropy(MaxEnt) is a supervised machine learning algorithm which is doesn't fall in to the category of data mining techniques.Although it's a statistical model which work on probability.Maximum Entropy or MaxEnt can act as a tool with vast field of data analysis which is somewhat linked to data mining.In simple words Maximum Entropy is like predicting the best option from the raw data which you have accumulated without adding extra assumed predictions.A probability can be estimated by amplifying the entropy of particular distribution. Twitter Data can be analzyed with help of Maximum Entropy.In Maximum Entropy Data is been cleaned by changing all text in lower cases to ensure that consistency is maintained within data.After Data Cleaning feature extraction and classification is done where certain words and their importance is weighed the words which occurs most are given higher weights and classification is perform based on it.Maximum Entropy can be used for sentiment analysis too as the classifier is trained with help of labeled datasets and new text data is been tested or classified with help of prior training and can categorized the data based on sentiments.Maximum Entropy can help discover meaningful insights through public opinion and sentiments.

Apriori Algorithm is an unsupervised algorithm which falls under category of data mining technique and specifically used for association rule mining.Apriori Algorithm is use to calculate the association rule and relation between two or more objects.The association rule defines how objects are related or associated to each other.The best example to understand apriori algorithm is if suppose you went to a online shopping site and decided to buy a earphone and after buying earphone you will also get suggestion to buy earphone pouch which help increases their sales.According to references we will see how apriori algorithm is used to mine data from facebook and other social platform and how that data is classified and it is secured.Data is collected from various posts,stories,content and trends in facebook.The data which was collected is then refined to minimize error while analysis.An ItemSet is generated which consist of frequently occurred on facebook.Association rules were applied on generated itemset Eg(Beautiful Flowers the word beautiful is been associated with flowers.This analysis help what topics and contents are trending on social media and then classification is done. This Algorithm helps discover frequent pattern and association which can help one to gain business insights more.

Neuro-Fuzzy Algorithm is combination of both supervised as well as unsupervised learning algorithm.The Neuro-Fuzzy Algorithm implements the logic of both Artificial Neural Network(ANN) and Fuzzy logic. As ANN are best

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306225|

suited to handle complex functionalities and architectures.Whereas Fuzzy Algorithm can control noisy and inconsistency in data.Data is been gathered from various social platforms and further it is send for cleaning process.Text Data is been cleaned to remove noise and inconsestencies present in it.Using Neuro-Fuzzy Algorithm the data is been split into different clusters for analysis purpose Advanced Apriori Algorithm is also used along with Neuro-Fuzzy Algorithm to analyze pattern and sentiment to discover more insights in it.Due to its supervised and unsupervised learning algorithm it can work well in classification,clustering and prediction but in terms of accuracy it cannot give much better accuracy as compared to SVM due to its accurate separation of classes through hyperplane.

Hidden Markov Model is a type of statistical model.HMM does not fall under the category of data mining techniques. It can act as a tool for data mining .HMM is widely used for modeling data which in sequential order and time series analysis.HMM describes probabilistic relationship among sequential events which are observed and sequential hidden states.The term “Hidden” refers as it is useful in scenarios when a system proudces data which is observed it is hidden or not known.Direct Access is not available to the internal state that produces outcome.HMM can be used as a tool to analyze behaviour of user and how user interacts with other on social media.This model can predict future on the basis of past data . Similarly,Data is been collected from social media and refined so that it can be easily analyzed by HMM.HMM identify hidden states and predicts future behaviour of user on past data.it also observe how user behaviour transist between each state of time.HMM can discover deep insights and help gains value in business and marketing on social media platforms.

III. CONCLUSION

According to our analysis and researches we saw all different types of data mining techniques that are used in social media .Through our research we figured out the best optimal and least optimal data mining technique that are used in social media.SVM fits as the best optimal data mining technique in social media for classification purpose.As it has the highest accuracy about(95-98%) .Due to its ability to handle vast amount of dataset without any slowdown it holds an upper edge among all the data mining techniques.Naive Bayes classifier has an accuracy of (75-80%) and has less optimality and accuracy than SVM.Decision Tree Classifier can be useful for classification and is not widely used in social media mining .K-Means Clustering is valuable for clustering data points having similar properties and to discover new patterns but has less precision as supervised learning. LDA can be useful for user behaviour prediction. Maximum Entropy and Apriori Algorithm can be used for specialized topics like sentiment analysis an d association mining.Neuro Fuzzy works very well with noisy data but lacks accuracy as compared to other data mining techniques.HMM is only valuable for specific data like sequential data and time series analysis. In summary,SVM is the most optimal and best technique.While Naïve Bayes and LDA can be used for specialization.Neuro Fuzzy,K-Means and Decision Tree are less optimal for data mining.

REFERENCES

1. David Ediger Karl Jiang et.al“Massive Social Network Analysis Mining Twitter for Social Good”, ICPP 2010
2. G Nandi, A Das, “A Survey on Using Data Mining Techniques for Online Social Network Analysis”, 2013.
3. Xi Long, Wenjing Yin, et.al “Churn Analysis of Online Social Network Users Using Data Mining Techniques”, IJEMS 2012.
4. Meenu Sharma, “Clustering In Data Mining : A Brief Review”, (IJCEM) Volume 1, Issue 5,August 2014.
5. Md. Ansarul Haque1, et.al “SENTIMENT ANALYSIS BY USING FUZZY LOGIC”, (IJCSEIT), 2014.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details