International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

A Recommendation System & their Performance Metrics using several ML Algorithms

Gattu Vijaya Kumar, Prasanta Kumar Sahoo, K.Eswaran

Abstract*: Recommendation systems are subdivision of Refine Data that request to anticipate ranking or liking a user would give to an item. Recommended systems produce user customized exhortations for product or service. Recommended systems are used in different services like Google Search Engine, YouTube, Gmail and also Product recommendation service on any E-Commerce website. These systems usually depends on content based approach. in this paper, we develop these type recommended systems by using several algorithms like K-Nearest neighbors(KNN), Support-Vector Machine(SVM), Logistic Regression(LR), MultinomialNB(MNB),and Multi-layer Perception(MLP). These will predict nearest categories from the News Category Data, among these categories we will recommend the most common sentence to a user and we analyze the performance metrics. This approach is tested on News Category Data set. This** data set having more or less 200k Headlines of News and 41 classes, ***collected from the Huff post from the year of 2012-2018.

Keywords*: Recommendation system: Support-Vector Machine: Multilayer Perceptron: K-Nearest neighbors: Logistic Regression: MultinomialNB.

I. INTRODUCTION

Recommended systems deals with recommending of products or items to a user based on their interest. The main reason we need a recommendation system in the current generation is because humans have extremely many alternatives to utilize required information which is popular from the Internet. The Recommend system solves the information overloading problem [1]. It is running based on three phase's object-data collection, similarity decision and prediction computation on the report of Chen and Anne Yun-An [2]. Mainly three types of recommendation systems are available they are Content-Based Recommendation system, Collaborative Filtering-Based Recommendation System and Hybrid Recommendation system which works effectively in the Media and Entertainment production. Data set used in this paper is News Category. It consist more or less 200k Headlines of News and 41 classes, the data set is Collected from Huff Post website from the year of 2012-2018. The below Figure shows the category wise data distribution.

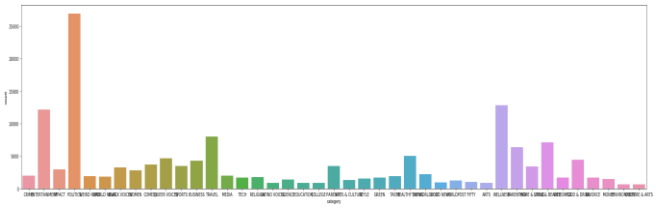


Fig 1. Flowchart of category wise data.

In this dataset total number of authors or 27135 members. Average News articles written by each author is 5.

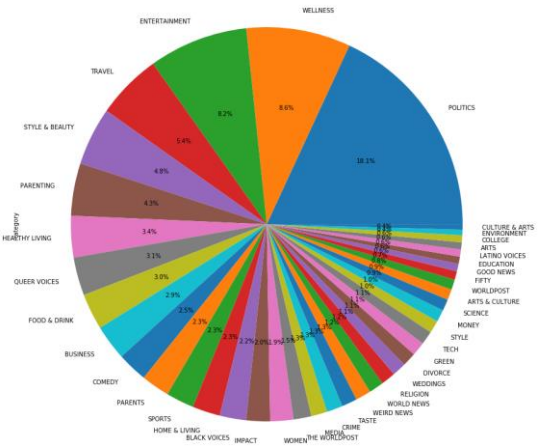


Fig 2. Flowchart shows contributions of authors on each category.

The above Fig2 shows that almost 35% of the news categories from Politics, Wellness and Entertainment.
Machine learning (ML) is promptly one of the vigorous methodology in present era. Machine learning (ML) is the analytical models and research-based study of algorithms that computer machine is use to perform particular task without being explicitly programmed.These algorithms used in variety of applications like Spam Filtering, Face Recognition, Speech Recognition and Computer Vision etc.
Machine learning algorithms are three types.

Revised Manuscript Received on February 05, 2020.
* Correspondence Author
Gattu Vijaya Kumar*, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India. E-mail:- vijaykumargattu1@gmail.com
Prasanta Kumar Sahoo, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India. E-mail- prasantakumars@sreenidhi.edu.in.
K.Eswaran, Department of Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana, India, E-mail- kumar.alpes@gmail.com

Supervised learning: In supervised learning the machine is learned from the data which have labels and tag values. By using labeled data we can easily predict the newly entered data. The supervised learning algorithm is similar to the students which are learning under supervision of teachers.

Unsupervised learning: Unsupervised learning is the machine is learned from the data which does not contain any labels or tag values. In unsupervised learning we classify or group the data by observing the similarity or relationship between the other data.

Reinforcement learning: The reinforcement learning algorithm is a one type of algorithm in which the machine is interacting with its environment by performing some actions and analyzes the data.

Algorithms

In this paper generally we used classification algorithms like Support-Vector Machine, K-Nearest Neighbors, Logistic Regression, Multinomial Naïve Bayes, Multilayer Perceptron to classify Data and we evaluated these algorithms performance metrics to compare which algorithm is given better results.

II. RELATED WORK

Xiwang Yang et.al [1] did online social voting using collaborative filtering based recommendation system. From the experiment with actual data, the group associated information can greatly increase the perfection of popularity-based voting recommendation, in general for cold users compare to social network information. In this Recommendation system of online social voting, used collaborative filtering algorithms like Matrix Factorization and Nearest Neighbor Method. The popularity based social voting recommendation leads to information overloading problem. Parteek Parhi et.al [2] conducted survey on technics of collaborative filtering methods. By using Neighborhood approach it's better to do recommendation on smaller dataset, whereas using Latent Factor approach the main advantage is user control of higher level. Neighborhood based approach takes matrix as input and predicts the model using comparability limitations. In Latent Factor approach from the user-rating matrix it's calculates the hidden characteristics and predicts the model. The both approaches Neighborhood and Latent Factor having problem of cold start and data sparsity. Paolo Cremonesi et.al [3] calculates the Top-n Recommendation tasks performance by using collaborative filtering algorithms like Non- Personalized models, Neighborhood models, Singular value decomposition and Latent factor models. This experiment done on two dataset Movie lens and Netflix datasets. Performance of Top-n recommendation system can be evaluated by using precision/recall and error metrics Mean Absolute Error and Root Mean Square Error. Depend on this tasks the error metrics Root Mean Square Error can work as quality representative. Xiwang Yang et.al [4] using social network information did Top-k Recommendation. Community - based Recommendation is expansive in the actuality, but recommendation of top-k using online community - based networks has subsist not adequate. He did inclusive research on raise the performance of top-k recommendation using certainty data obtain from the online community - based networks. Through experiments on Flixster and Epinions dataset find one important thing by using Matrix Factorization model and Nearest Neighborhood models certainty data improves the top-k recommendation. From the study of all the above work we are going to implement A Recommendation system and Evaluates the performance metrics analysis by using several Machine Learning Algorithms.

III. PROPOSED WORK

A. *Support-Vector Machine*

Support-Vector Machine is a supervised learning algorithm, used for classification and regression analysis. In Support-Vector Machine algorithm each data entity plot as a point in n- dimensional volume accompanied by the value of every peculiar actuality of the value of a specific coordinate. The classification is done by using Hyper-plane/Line which will segregate the two classes. It's classifies both linear and non-linear problems using different kernels.

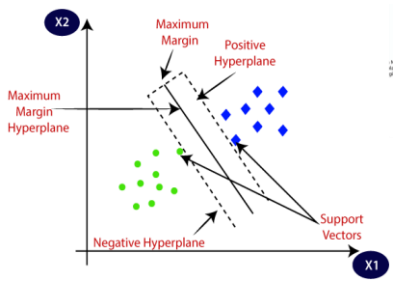


Fig 3. General classification of hyper plane.

B. *K-Nearest neighbors*

The K-Nearest neighbor is extremely easy to classify the data and it is also be used in regression problems like SVM. It is also supervised learning algorithm because, the model is trained and evaluated by class labels. KNN uses the similarity features to predict the nearest neighbors. To discover K-Nearest neighbors, K value will be used where, K is an integer value. Based on "K" value it calculate the Euclidean distance between data points.

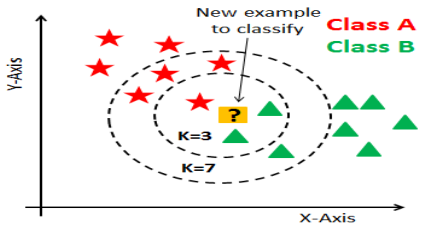


Fig 4. K-NN classification based on Euclidian Distance.

C. *Multinomial Naïve Bayes*

Multinomial Naïve is a simple probabilistic model its purely depends on Bayes theorem. Bayes theorem is used to solve conditional probability problems. Multinomial Naïve Bayes specifically used for Natural Language processing problems because, it is fast and reliable when compare to similar classification algorithms like SVM and Neural Network.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Fig 5. Bayes Theorem.

D. *Logistic Regression*

In Machine learning the Logistic algorithm is also one of the classification analysis algorithm. It will predict the only specific values, according to particular sequence to map predicted values to probabilities used sigmoid function. It also same like linear regression but whereas linear regression deals with only binary class. In Logistic Regression there are three types. Binary Logistic Regression, Multi-class Logistic Regression and Ordinal Logistic Regression. Multi-class Logistic Regression deals with other than two class.

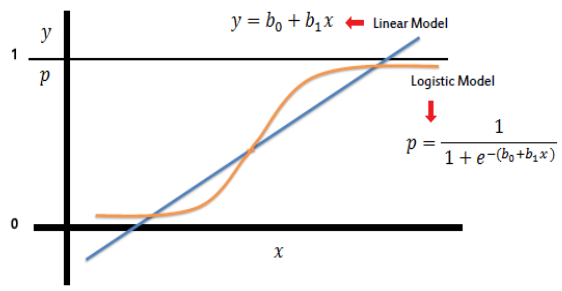


Fig 6. General classification of Logistic Regression.

E. *Multilayer-Perceptron*

In Machine learning Perceptron is Artificial Neural Network entity, it is also a supervised learning algorithm for binary classifiers. Linear separable patterns or done by single layer perceptron. Perceptron's are two types' Single layer and Multilayer perceptron.

In Multilayer Perceptron more than three layers are available like input layer, hidden layer and output layer, because of these layers it have the appreciable processing power. It will take inputs and those weights do some process and produce the output to the output layer.

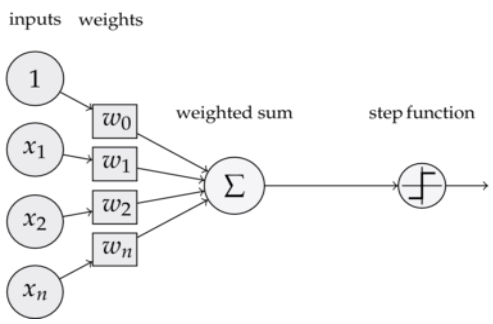


Fig 7. Multilayer-Perceptron

In proposed work we use above classifiers to classify the News headlines category from the dataset, build a recommendation system and evaluate the performance of those algorithms.

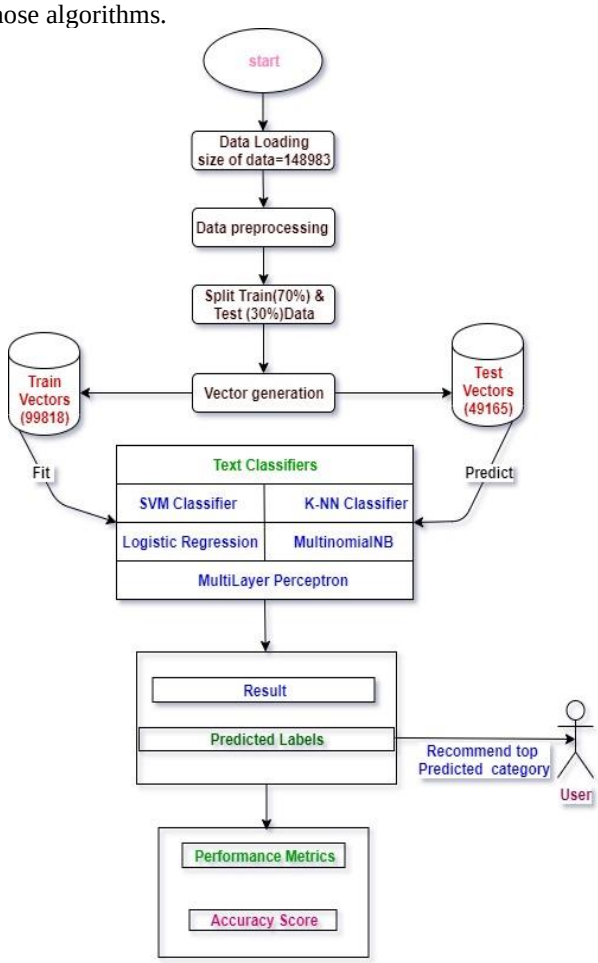


Fig 8: Architectural prototype of proposed system.

As signify in the above architectural prototype diagram, it's involves the proposed prototype of Recommendation system and their performance metrics analysis by using several Machine learning algorithms.

1. The foremost step in the structure of Recommendation system is sort out the preferred News Category dataset for Text classification. The Data contains Headline of the news, short Description of the news and categories of news and etc.
2. After data set is collected, we need to extract the features for that, apply the preprocessing technique to dataset using Natural Language Tool Kit (NLTK).

It will divide the Text sentence into meaningful tokens, remove stop words and convert all characters into lowercase and remove empty cells & clean the unwanted data from the dataset.

3. Once the preprocessing of headline and short description of news is done. Split Train Data and Test Data in required percentage.
4. The result of train and test split data given to TFDIF Vectorizer. This will convert the text data into vectors by using (fit transform) model from the TFIDF vectorised. These Vector stores data in the grouping of sparse matrix. The matrix incorporate with dominant amount of Zeros is called sparse matrix. The foremost advantage with sparse matrix is consumes less memory because it contains less number of non-zero elements.

5. Fit the Train vectors to different type's classification algorithm like Support-vector Machine, K-Nearest Neighbors, Logistic Regression, Multinomial Naïve Bayes, and Multilayer Perceptron. After completion of data training, We Test model with test vectors and predict the labels.

6. calculating the accuracy score of various classifiers using confusion matrix of the each classifier and differentiate the accuracy score of different algorithms to see which algorithm provides better result.

7. To recommend n -nearest categories of a test sentence we calculated weights of each and every model and Multiplied weights with actual train & actual test data.

8. After multiplication of weights with train and test data we got another train and test data. Fit the train data to nearest neighbors and test the model with test data it will predict the n- nearest categories to a specific test point. Here n will returns the number of nearest points to a single test point

IV. PERFORMANCE METRICS ANALYSIS

In The Machine learning assessing algorithms is extremely important segment of the project. Classification Machine learning models we use Accuracy score, Precision, Recall and F1 score.

Confusion Matrix

		Actual (Positive)	Actual (Negative)
Predicted	(Positive)	True Positive(TP)	False Positive(FP)
	(Negative)	False Negative(FN)	True Negative(TN)

Accuracy score represent as the amount of exact predictions divided by the total amount of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

V. COMPARISON TABLE

From the comparision of the below table Support-Vector Machines provides 75.13% highest accuray.

Classifiers	K-Value	Train Data size	Test Data size	Tfidf with Original Data
SVM	-	99818	49165	75.13%
Logistic Regression	-	99818	49165	73.66%
Multilayer Perceptron	-	99818	49165	73.48%
Multinomial Naive Bayes	-	99818	49165	68.59%
K-NN	9	99818	49165	62.61%
K-NN	5	99818	49165	54.87%
K-NN	3	99818	49165	54.72%
K-NN	1	99818	49165	52.83%,

Table 1. This table of content shows Accuracy of the proposed Recommendation system using Different classification algorithms for News Category Dataset.

VI. RESULTS OF N- NEAREST RECOMMENDATION

1) N-Recommendations using Logistic Regression.

Test_sentence1: '10 Years Ago, These People Tried To Drive Undocumented Immigrants Out Of Town. Now, They're Advising Trump.They could push for harsh new immigration restrictions – on a national scale.'Dana Liebelson'

Test_Category: 'POLITICS'

Nearest_Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence2: 'Sunday RoundupThis week the presidential circus rolled into Wisconsin, with Ted Cruz scoring a big victory against a stumbling Trump. New York Magazine\'s Gabriel Sherman suggested those wobbles might be due to burnout. "People who know Trump say they\'ve never seen him so tired," he wrote. Supporting evidence comes from modern science, as sleep deprivation symptoms include lack of judgment, inability to process basic information, irritability, mood swings and a paranoid tendency to spout conspiracy theories. Sound familiar? Trump\'s biggest contribution to the discourse might be as a cautionary tale about exhaustion, and the resulting consequences of lack of impulse control. And as the campaign heads to the city that never sleeps, I\'ll be heading west on a different campaign, to change the way we think about sleep -- that includes a college tour and a campaign against drowsy driving with Uber. In the meantime, all the candidates could do the electorate a favor and get some sleep.''Arianna Huffington, Contributor

Test_Category: 'POLITICS'

Nearest Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence3: "Why We Can't Shut Up About Beauty and Praise Our Girls for Being SmartI totally get the impulse to want to shun the topic, because the topic brings a lot of people a lot of pain. But, not talking about it doesn't prepare our girls for reality."'Amanda King, Contributor\nBlogger, Last Mom On Earth'

Test_Category:'PARENTING'

Nearest_Points: 'PARENTING' 'PARENTING' 'PARENTING'

2) N-Recommendations using Logistic Regression.

Test_sentence1: '10 Years Ago, These People Tried To Drive Undocumented Immigrants Out Of Town. Now, They're Advising Trump.They could push for harsh new immigration restrictions – on a national scale.'Dana Liebelson'

Test_Category: 'POLITICS'

Nearest_Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence2: 'Sunday RoundupThis week the presidential circus rolled into Wisconsin, with Ted Cruz scoring a big victory against a stumbling Trump. New York Magazine\'s Gabriel Sherman suggested those wobbles might be due to burnout. "People who know Trump say they\'ve never seen him so tired," he wrote. Supporting evidence comes from modern science, as sleep deprivation symptoms include lack of judgment, inability to process basic information, irritability, mood swings and a paranoid tendency to spout conspiracy theories. Sound familiar? Trump\'s biggest contribution to the discourse might be as a cautionary tale about exhaustion, and the resulting consequences of lack of impulse control. And as the campaign heads to the city that never sleeps, I\'ll be heading west on a different campaign, to change the way we think about sleep -- that includes a college tour and a campaign against drowsy driving with Uber. In the meantime, all the candidates could do the electorate a favor and get some sleep.''Arianna Huffington, Contributor

Test_Category: 'POLITICS'

Nearest Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence3: "Why We Can't Shut Up About Beauty and Praise Our Girls for Being SmartI totally get the impulse to want to shun the topic, because the topic brings a lot of people a lot of pain. But, not talking about it doesn't prepare our girls for reality."'Amanda King, Contributor\nBlogger, Last Mom On Earth'

Test_Category:'PARENTING'

Nearest_Points: 'PARENTING' 'PARENTING' 'PARENTING'

3) N-Recommendations using Multilayer Perceptron.

Test_sentence1: '10 Years Ago, These People Tried To Drive Undocumented Immigrants Out Of Town. Now, They're Advising Trump.They could push for harsh new immigration restrictions – on a national scale.'Dana Liebelson'

Test_Category: 'POLITICS'

Nearest_Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence2: 'Sunday RoundupThis week the presidential circus rolled into Wisconsin, with Ted Cruz scoring a big victory against a stumbling Trump. New York Magazine\'s Gabriel Sherman suggested those wobbles might be due to burnout. "People who know Trump say they\'ve never seen him so tired," he wrote. Supporting evidence comes from modern science, as sleep deprivation symptoms include lack of judgment, inability to process basic information, irritability, mood swings and a paranoid tendency to spout conspiracy theories. Sound familiar? Trump\'s biggest contribution to the discourse might be as a cautionary tale about exhaustion, and the resulting consequences of lack of impulse control. And as the campaign heads to the city that never sleeps, I\'ll be heading west on a different campaign, to change the way we think about sleep -- that includes a college tour and a campaign against drowsy driving with Uber. In the meantime, all the candidates could do the electorate a favor and get some sleep.''Arianna Huffington, Contributor

Test_Category: 'POLITICS'

Nearest Points: 'POLITICS' 'POLITICS' 'POLITICS'

Test_sentence3: "Why We Can't Shut Up About Beauty and Praise Our Girls for Being SmartI totally get the impulse to want to shun the topic, because the topic brings a lot of people a lot of pain. But, not talking about it doesn't prepare our girls for reality."'Amanda King, Contributor\nBlogger, Last Mom On Earth'

Test_Category:'PARENTING'

Nearest_Points: 'PARENTING' 'PARENTING' 'PARENTING'

4) N-Recommendations using Multinomial Naive Bayes.

Test_sentence1: "The Style Selfie Rears Its Head at the Emmy'sThis photo is not usually taken by the subject, but rather by a friend (or even stranger) but preferably by a friend who can continue to take photos until you run out of storage or until the perfect editable image surfaces."'Alisa Wolfson, Contributor\nContributor'

Test_Category: 'STYLE & BEAUTY''

Nearest_Points: 'STYLE & BEAUTY' 'ENTERTAINMENT' 'STYLE & BEAUTY''

Test_sentence2: "Why We Can't Shut Up About Beauty and Praise Our Girls for Being SmartI totally get the impulse to want to shun the topic, because the topic brings a lot of people a lot of pain. But, not talking about it doesn't prepare our girls for reality."'Amanda King, Contributor\nBlogger, Last Mom On Earth'

Test_Category: 'PARENTING'

Nearest Points: 'PARENTS' 'PARENTS' 'PARENTING'

Test_sentence3: 'From America\'s Next Top Model to Green AdvocateAfter receiving the title "Hippie Hannah" from none other than the queen of America\'s Next Top Model, Tyra Banks, Hannah has embraced her nickname for what it represents. She passionately promotes global awareness and sustainable lifestyle practices through food, fitness and the arts.'
'Nova Lorraine, Contributor\nEditor in Chief, Raine Magazine'

Test_Category:''STYLE & BEAUTY''

Nearest_Points: 'STYLE & BEAUTY' 'ENTERTAINMENT' 'STYLE & BEAUTY'

5) N-Recommendations using K-Nearest Neighbors.

Test_sentence1:'From America\'s Next Top Model to Green AdvocateAfter receiving the title "Hippie Hannah" from none other than the queen of America\'s Next Top Model, Tyra Banks, Hannah has embraced her nickname for what it represents. She passionately promotes global awareness and sustainable lifestyle practices through food, fitness and the arts.'
'Nova Lorraine, Contributor\nEditor in Chief, Raine Magazine'

Test_Category: 'STYLE & BEAUTY'

Nearest_Points: 'QUEER VOICES' 'STYLE & BEAUTY' 'STYLE & BEAUTY'

Test_sentence2: "The Style Selfie Rears Its Head at the Emmy'sThis photo is not usually taken by the subject, but rather by a friend (or even stranger) but preferably by a friend who can continue to take photos until you run out of storage or until the perfect editable image surfaces."'Alisa Wolfson, Contributor\nContributor'

Test_Category: 'STYLE & BEAUTY'

Nearest Points: 'WELLNESS' 'WEDDINGS' 'ENTERTAINMENT'

Test_sentence3: "Why We Can't Shut Up About Beauty and Praise Our Girls for Being SmartI totally get the impulse to want to shun the topic, because the topic brings a lot of people a lot of pain. But, not talking about it doesn't prepare our girls for reality."'Amanda King, Contributor\nBlogger, Last Mom On Earth'

Test_Category:''PARENTING'

Nearest_Points:'WOMEN' 'POLITICS' 'POLITICS'

VII. CONCLUSION

As shown in all the above results, this paper use five classifiers are Support-Vector Machines, Logistic Regression Multinomial Naive Bayes, Multilayer Perceptron and K-Nearest Neighbors .It was observed that from the comparison of all the algorithms Support-Vector Machine gives 75.13% accuracy. From the News category dataset categorizes the category of sentence and recommended these categories to user. In future we try to improve the model using different word or embedding sentences victimizers.

REFERENCES

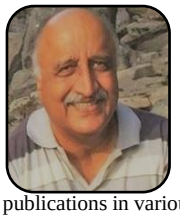
1. Xiwang Yang et al., "Collaborative Filtering-Based Recommendation of Online Social Voting", (IEEE) TRANSACTIONS ON COMPUTAIONAL SOCIAL SYSTEMS, 2017.
2. Parteek Parhi et al., "A Survey of Methods of Collaborative Filtering Techniques."(IEEE) International conference on Inventive Systems and Control (ICISC-2017).
3. Paolo Cremonesi et al., "Performance of Recommender Algorithms on Top-N Recommendation Tasks", ACM -2010.
4. Xiwang yang et al., "On Top-K Recommendation using Social Networks", 2012 ACM 978-1-4503-1270-7/12/09.
5. Yehuda Koren, "Collaborative Filtering with Temporal Dynamics", COMMUNICATIONS OF THE ACM, DOI:10.1145/1721654,1721577.
6. H.Gao et al., "Content-aware point of interest recommendation on location-based social networks," 2015.
7. G. Adomavicius et al., "Toward the next generation of recommender system: A Survey of state of the art and possible extensions," IEEE vol.17, Jun 2005.
8. Jie Qin et al., "collaborative filtering recommendation algorithm based on weighted item category," IEEE 2016 Conference.
9. Mukesh Kumar et.al., "Item-Based Collaborative Filtering in Movie Recommendation in Real Time" IEEE 2018 Conference.

AUTHORS PROFILE

Gattu Vijaya Kumar Received Bachelor's Degree in Computer Science & Engineering from Rajiv Gandhi University Of Knowledge Technologies(RGUKT), Basar, Adilabad, India, in 2016. Pursuing Masters from Sreenidhi Institute of Science and Technology and presently doing Research on Machine Learning.

Dr. Prasanta Kumar Sahoo Professor, Department of Computer Science and Engineering, Sreenidhi Institute of Science & Technology, Hyderabad. He completed his Ph.D. from Fakir Mohan University, Odishain Computer Science Engineering. He has 17 years of teaching, research and administrative experience. He has earlier worked as Head of the Dept. in both CSE and IT dept. in various reputed Engineering Colleges. His Research Interest includes Cyber Security, Information Security and Data Mining. He has published around 50 research papers in various reputed journals both at national and International level. His research papers were cited both at national and international level, so far by 41 citation and 1567 reads as per Google Scholar and research Gate report.

Many times Dr. Prasanta Kumar Sahoo won the best teacher award in various colleges for his contribution to the teaching and learning process. He is Certified Professional from BalaBit, completed Electronic Contextual Security Intelligence exam Intermediate Level (ECSI). He has guided more than 50 projects both at UG and PG level. He has delivered more than 15 guest lecturers. He has organized three national conference and nine faculty development program with an immense success.



K. Eswaran joined SNIST from 1999 after leaving BHEL R&D as Additional General Manager. He has Masters and Ph.D., degrees from IIT Kanpur and University of Madras respectively, He now works in the area of Artificial Intelligence involving Neural Networks and Image Processing using Pattern recognition methods. He has more than 40 publications in various conferences. Many of his publications are referred to by present researchers, even after several decades. He has also won several best teacher awards the latest beings as Best Faculty Award 2013-2014, Computer Science from Cognizant for the south India Area.