ISSN No. 2581 - 5571
RNI NO. UPENG/2017/72946

LAKSH
Estd. 1996

IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH

VOL. 5 DECEMBER 2020

The Annual Refereed Journal of the Computer Applications of the Institute of Professional Excellence and Management

Published by :

Dept. of Computer Applications
Institute of Professional Excellence and Management

(ISO 9001:2015 Certified, NAAC Accredited)

A-13/1 South Side G.T Road, Industrial Area
NH-24 Bye Pass, Ghaziabad (U.P) -201010
Ph.: 0120-4174500, Fax: 0120-4174500
E-Mail:journalit@ipemgzb.ac.in
Website : www.ipemgzb.ac.in

RS. 300 (ANNUAL SUBSCRIPTION)

CONTENTS

Application of Machine Learning for Predictive Analytics: Indian Premier League (IPL) T-20 Cricket Matches
Subhashish Mahata, Neetu Kamra & Naina Kumari Agarwal

Power System Security Assessment using K-Nnearest Neighbour
P. Praveen & Dr. C.V.K. Bhanu

Semantic Information Retrieval:Unboxing the Complexity in Big Data
Khushbu A Patel & Dr. Mahammad Idrish I. Sandhi

Social Media Analytics Platform to assist Business Decision Making for Small and Medium Enterprise in Indonesia
Muhammad Arriandito Arya Saputra, Manahan Siallagan, Santi Novani, Lidia Mayangsari, Yogie Setiafriawan & Puteri Annisa Tsamrotul Fuadah

A review paper on Identification of the CAPTCHA with the advancement of Machine Learning Techniques
Surendra Kumar Pathak, Dr. Naveen Kr Singh & R.K. Maurya

Customer Behaviour Prediction using Propensity Model
Dr. Remya K Sasi & Hima John

Estimate of Modeling Units for Hindi Speech Recognition using Artificial Intelligence
Arjun Kumar, Achintya Kr. Pandey & Sudarshan Singh

Crypto currency: A Futuristic Digital Currency for the Digital World
Dr. Gavendra Singh, Mr. Afendi Abdi, Mr. RajaSekhar Boddu & Mr. Adugna Alemayehu

Blockchain with Artificial Intelligence
Shrikant Patel, Indu Jolly, Girish Kumar Sharma & Sanjay Kumar

Investment Analysis based on the Principles of Value Investing using Machine Learning
Tanay Mehendale, Anand Mane, Vishal Ramina & Shreyas Kailasnathan

The Future of Supply Chain – Data Logging Via Internet of Things (IoT)
Savita Singh & Abhijeet Kumar Singh

AI Assistant to support students and the users
Nandini Bagga, Pratikshit Vashistha, Palak Yadav & Dr. Arvind Kumar

IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH

The Annual Refereed Journal of Computer Applications of the Institute of Professional Excellence and Management

Vol. 5 December 2020

Founder, IPEM Group of Institutions

Dr. B.S. Goel

(04.08.1937-10.01.2017)

A Visionary, Educationist & Philanthropist with values

SHIKSHA BHARATI AWARD

presented to Dr. B.S. Goel Executive Director

INDIAN ACHIEVERS AWARD

Best Academic Excellence Institution in NCR

Editorial Board:

Printed and Published by Mr. Anupam Goel on behalf of Laksh Educational Society and Printed at Ghaziabad Offset Press, 133, East Model Town Tehasil Road, Ghaziabad and Published at Institute of Professional Excellence and Management, A-13/1, South Side G.T. Road Industrial Area, NH-24 Bypass Ghaziabad (U.P.) 201010 -INDIA. Editor: Dr. Sugandha Goel

From the Editorial Board

We are glad to present the Fourth Edition of the IPEM Group of Institutions, Computer Applications Journal "IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH", December 2020. However, we braved all the odds, and published the issue as always, on time. We followed a rigorous method to select the papers. All the papers we have included in this issue of IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH are peer reviewed and only those papers which went through this rigor have been given space in this Journal.

This Journal attempts to document and spark a debate on the research focused on technology in the context of emerging technologies. The area could range from Computational Intelligence, Cyber Security Challenges, Image Thresholding Techniques, Cloud based CRM etc. These technologies could be from very sophisticated to very elementary, but in term of impact they would be capable of being commercialized, scaled up and focused on real life challenges.

We sincerely hope that these in-depth research papers, focusing on different technologies, will further stimulate the academic research, and will help in developing an insight into the concerned areas. We are eagerly waiting for your critical response which we shall incorporate in the forthcoming issues. We are greatly indebted to the paper writers who took keen interest and submitted their research papers on time. It is because of the sincere efforts of these people that the IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH is in your hands today.

We are grateful to our Secretary - Mr. Anupam Goel who provided all the moral and financial support to publish the IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH.

IPEM JOURNAL OF COMPUTER APPLICATION & RESEARCH

The Annual Refereed Journal of Computer Applications of the Institute of Professional Excellence and Management

Vol. 5 December 2020

Contents

01 **Application of Machine Learning for Predictive Analytics: Indian Premier League (IPL) T-20 Cricket Matches** **02**
Subhashish Mahata, Neetu Kamra & Naina Kumari Agarwal

02 **Power System Security Assessment using K-Nnearest Neighbour** **12**
P. Praveen & Dr. C.V.K. Bhanu

03 **Sementic Information Retrieval:Unboxing the Complexity in Big Data** **21**
Khushbu A Patel & Dr. Mahammad Idrish I. Sandhi

04 **Social Media Analytics Platform to assist Business Decision Making for Small and Medium Enterprise in Indonesia** **29**
Muhammad Apriandito Arya Saputra, Manahan Siallagan, Santi Novani, Lidia Mayangsari, Yogie Setiafriawan & Puteri Annisa Tsamrotul Fuadah

05 **A review paper on Identification of the CAPTCHA with the advancement of Machine Learning Techniques** **34**
Surendra Kumar Pathak, Dr. Naveen Kr Singh & R.K. Maurya

06 **Customer Behaviour Prediction using Propensity Model** **38**
Dr. Remya K Sasi & Hima John

07 **Estimate of Modeling Units for Hindi Speech Recognition using Artificial Intelligence** **44**
Arjun Kumar, Achintya Kr. Pandey & Sudarshan Singh

08 **Crypto currency: A Futuristic Digital Currency for the Digital World** **48**
Dr. Gavendra Singh, Mr. Afendi Abdi, Mr. RajaSekhar Boddu & Mr. Adugna Alemayehu

09 **Blockchain with Artificial Intelligence** **53**
Shrikant Patel, Indu Jolly, Girish Kumar Sharma & Sanjay Kumar

10 **Investment Analysis based on the Principles of Value Investing using Machine Learning** **64**
Tanay Mehendale, Anand Mane, Vishal Ramina & Shreyas Kailasnathan

11 **The Future of Supply Chain – Data Logging Via Internet of Things (IoT)** **72**
Savita Singh & Abhijeet Kumar Singh

12 **AI Assistant to support students and the users** **78**
Nandini Bagga, Pratikshit Vashistha, Palak Yadav & Dr. Arvind Kumar

Application of Machine Learning for Predictive Analytics: Indian Premier League (IPL) T-20 Cricket Matches

Subhashish Mahata*
Neetu Kamra**
Naina Kumari Agarwal***

Introduction

Data is the ultimate Wealth in today's world. With the enhancement of technology, data has become the most powerful input in every sector. The approach of sports management also changed in a rapid fashion. Nowadays, sports managers and stakeholders are more emphasising on data for decision making.

Analytics is a hot topic in the arena of sports because of easy availability of huge live and historical data. Cutting edge technology like Machine Learning is being used in Sports Analytics. Sports analytics is a powerful method to extract knowledge from both live and historical sports data. Sports analytics in not only focus on prediction part, it also involves descriptive and predictive part. Sports analytics is now serving for each kind of sports.

Sports analytics is not new in cricket. Cricket being an attractive and profitable sports, there are so many stakeholders, involved in this sports. That's why decision making process is very critical in cricket. Analytics is a supporting pillar for every decision maker in this game.

Indian Premier League, known as IPL, is the most valuable and Expensive T20 cricket league in the world. It is the one of the biggest event for players, Team owners, business mans. IPL was started in 2008 by The Board of Control for Cricket in India (BCCI). The league is used to play in double round-robin league and playoffs format with 8 teams from different cities of India. The brand value of the IPL in 2019 was ₹475 billion (US$6.7 billion). According to BCCI, the 2015 IPL season contributed ₹11.5 billion (US$160 million) tothe GDP of the Indian economy.

Almost every IPL team's Management use Analytics for better games. Not only team's owners, there are several betting and fantasy cricket platform, which are highly depend upon analytics for their success. Analytics can help all of them for their success.

The research paper tries to predict the IPL matches using machine learning models with variables like match_id, inning, batting_team, bowling_team, over, ball, batsman, non_striker, bowler is_super_over, wide_runs, bye_runs, legbye_runs, noball_runs, penalty_runs, batsman_runs, extra_runs, total_runs, player_dismissed, dismissal_kind, fielder. To find the result it uses different models like Logistic Regression model, Naïve Bayes Model, Support Vector Machine(SVM) , Decision tree. The result of the study shows that for IPL game, Teams, Venue, Winning Toss, Venue of the Match and Decision after winning the toss are important influencers to win a match. Different Machine Learning helps to predict outcome of a match. Right selection of Machine Learning Model helps to increase Accuracy of Prediction. From Different Classification Models, Support Vector Machine, Decision Tree and Random forest are best to predict outcome of an IPL games. All of the following gives almost 88% accuracy Level. The study has been conducted from data of Kaggle. Secondary data has been used for the analysis.

*Student, Lloyd Business School
**Professor, Lloyd Business School
***Student, Lloyd Business School

Literature Review

In this there are many past research paper which has been discussed. The Predicting outcome of Indian Premier league using machine learning by Rabindra Lamsal, Ayesha Choudhary in 2018 in which there are variables utilized like home team, away team, toss winner, venue, umpires, home team score, away team score, power play score, playing 11 players, Number of wickets taken, Number of dot balls given, Number of fours, Number of sixes, Number of catches, Number of stampings. The Multilayer perception classifier outperformed other classifiers with correctly predicting 43 out of 60, 2018 Indian Premier League matches. The Twenty 20 format of cricket carries a lot of unpredictable, because a single over can change the continued pace of the match. Increased prediction accuracy in the game of cricket using machine learning by Kalpdrum Passi, Nirav kumar Pandey in 2018 in which variables was No. of Innings, Batting Average, Strike Rate, Highest Score, Overs, Bowling Average, Bowling Strike Rate, Four/Five Wicket Haul, Venue ,Centuries, Fifties, Batting, Match Time, Hand, Match Type, Batting Position, Bowling Hand in which analysis was Random Forest builds the most accurate prediction models for both batting and bowling in all the cases. Also, the accuracy of the models increases as increase the size of the training data set for all increase the size of the training data set for all algorithms except in case of Naïve Bayes for batting where the accuracy decreases as we increase the dimensions of the training set. Selection of the proper players for every match plays a big role in a team's victory. An accurate prediction of what percentage runs a batsman is probably going to attain and the way many wickets a bowler is probably going to require match which will help the team to select the players for a specific match. Prediction of Live Cricket Score and Winning by Pramila M. Chawan in 2018 in which variables used was Pitch, Toss, Team strength, Home Ground Advantage in which result

was a predictive model, a user makes a prediction on every game, and ends up watching that game to check if his prediction is going right. Thus the project will not only improve the existing system of Fantasy Cricket, but will also augment the reach of Cricket in India. Cricket Analytics and Predictor by Mr. Suyash Mahajan, Ms. Gunjan Kandhari, Ms. Salma Shaikh, Ms. Rutuja Pawar, Mr. Jash Vora, Ms. A. R. Deshpande in 2019 in which variable used was City, Venue, Toss Result, Home Team, Away Team in which the previous data, it is beneficial to the owner to get the details of the IPL match played and the users who predict the winning percentage of the team and get the statistics of the player.

Sl. No	Name of the Paper	Author(s)	Year	Variable Used	Result
1	Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning	Rabindra Lamsal, Ayesha Choudhary	2018	home team, away team, toss winner, venue, umpires, home team score, away team score, power play score, playing 11 players, Number of wickets taken Number of dot balls given, Number of fours, Number of sixes, Number of catches, Number of stampings.	The Multilayer perception classifier outperformed other classifiers with correctly predicting 43 out of 60, 2018 Indian Premier League matches. The Twenty20 format of cricket carries a lot of randomness, because a single over can completely change the ongoing pace of the game. Indian Premier League is still at infantry stage, it is just a decade old league and has way less number of matches compared to test and one-day international formats.

2	Increased prediction accuracy in the game of cricketusing machine learning	Kalpdrum Passi, Niravkumar Pandey	2018	No. of Innings, Batting Average,Strike Rate,Highes Score,Overs,Bowling Average,Bowling Strike Rate,Four/Five Wicket Haul,Venue,Centuries,Fifties,Batting,Match Time,Hand,Match Type,Batting Position,Bowling Hand	Random Forest builds the most accurate prediction models for both batting and bowling in all the cases. Also, the accuracy of the models increases as increase the size of the training dataset for all algorithms except in case of Naïve Bayes for batting where the accuracy decreases as we increase the size of the training set. Selection of the right players for each match plays a significant role in a team's victory. An accurate prediction of how many runs a batsman is likely to score and how many wickets a bowler is likely to take in a match will help the team management select best players for each match.
3	Prediction of Live Cricket Score and Winning	Pramila M. Chawan	2018	Pitch,Toss,Team strength,Home Ground Advantage	a predictive model, a user makes a prediction on every game, and ends up watching that game to check if his prediction is going right Thus the project will not only improve the existing system of Fantasy Cricket, but will also augment the reach of Cricket in India
4	Cricket Analytics and Predictor	Mr. Suyash Mahajan, Ms. Gunjan Kandhari, Ms. Salma Shaikh, Ms. Rutuja Pawar,Mr. Jash Vora,Ms. A. R. Deshpande	2019	City, Venue, Toss Result, Home Team, Away Team	From the previous data, it is beneficial to the owner to get the details of the IPL match played and the users who predict the winning percentage of the team and get the statistics of the player.

Research Methodology

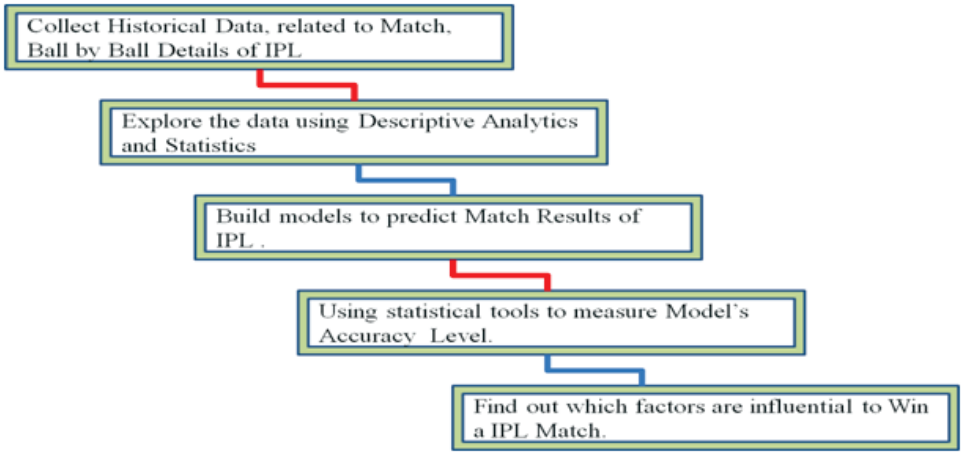
Research Objective: The objective of this research is to observe impact of different Machine learning models in Prediction of an IPL match. Another objective of this study is to explore information, pattern related to Matches, Player etc. using descriptive analysis so as to increase the decision making effectiveness.

Purpose of Study: Betting is an illegal activity some county like India, but most of the country, it's a profitable business because Betting is not only skill of gambling, it is also a game of statistical skills. Not only betting , there are many fantasy match platform like Dream11,My cricket circle where millions of users invest money to get handsome amount of profits where the statistical skills and data is important.

This study primarily aims to find out different statistical measures from IPL historical data and predict outcome of a match based on important factors to help users of betting sites and fantasy cricket league with scientific proof to support in their decision making process.

Methodology

Overall Project Plan (Context Diagram):



Sample Design:

Secondary data is what is collected by someone other than the Researchers. Some Common sources of secondary data include government public services department's Repository, libraries, internet searches and censuses.

For this project, researchers have used Secondary data source to collect data.

Data Source:

For this project work, data has been taken from Kaggle.com. Kaggle is subsidiary of Google LLC. It is an online community of data scientists and machine learning aspirant. It is also a repository of open source data.

Analytical Methodology:

This Project work focus on following two Analytical Methods –

I. Descriptive Analytics
II. Predictive Analytics

A. **Descriptive Analytics:** Descriptive Analytics is a Method use in primary stages of any Analytics project to create a summary of historical data to mine useful knowledge, based on which further analysis can be done. In simple language, Descriptive analytics answered question like "what happened?" .

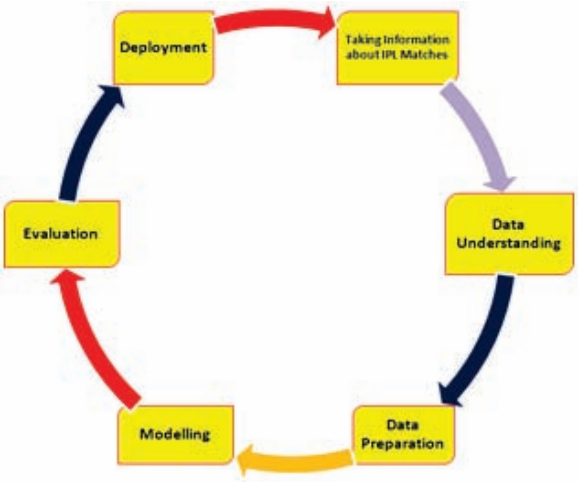
In our Project work, Descriptive model focus on two aspects:

i. Describe the data statistically.
ii. Describe important factors.

B. **Predictive Analytics:** Predictive Analytics is a Method use in Advance stages of Analytics Projects to Predict Unknown future events based on different factors. Predictive Analytics use different Algorithms to build predictive models. Some of popular Algorithms, used in Predictive Modelling, are – Logistic Regression, Linear Regression, KNN, Decision Tree, Random Forest, SVM etc.

For our project work, we will use predictive analytics to predict result of any matches.

Data Flow Diagram (DFD):



Tool Used

A. **IBM Cognos:** Cognos is a web-based Business Intelligence platform by IBM. It provides a toolset for analytics, reporting, and monitoring of different metrics. The IBM Cognos consists of different components which is used to meet the different information requirements in any company. IBM Cognos has components such as IBM Cognos Framework Manager, IBM Cognos Cube Designer, IBM Cognos Transformer to help to analyse data easily. Cognos will use in this project work as a Descriptive analytics tool.

B. **Tableau:** Tableau is an American interactive data visualization software . User has many advantages of using tableau as the software can handle different format of data, provide attractive Visualizations. Tableau will use as a visualization tool for this project report.

C. **Python:** Python is an interpreted, high-level, general-purpose language. Python is developed by Guido van Rossum in 1991.Its language uses object-oriented approach which help programmers to write clearer and logical code for any type of projects. Because of its extensive libraries, Great Community, memory management ,python is very popular among Machine Learning community. For this Project report , Python will be used for both descriptive and predictive analytics.

D. **IBM Watson Studio (Watson Assistant):** IBM Watson Studio is an software to make it easy to develop, train, manage models, and deploy AI applications and it used intent,entities and dialog. It is evolving with lot of new features to build Artificial Intelligence applications. IBM Watson Assistant is a cloud service that allows to develop virtual assistantin the software they are developing and brand the assistant as their own.Watson assistant will help to develop a virtual assistant for this project work.

Analysis and Interpretations

A. Descriptive Analysis

Total Match, Session and Different teams, Venue

```
#Total Matches
matches['id'].count()
```

```
756
```

```
In [83]: #Total Session
         len(matches['season'].unique())

Out[83]: 12
```

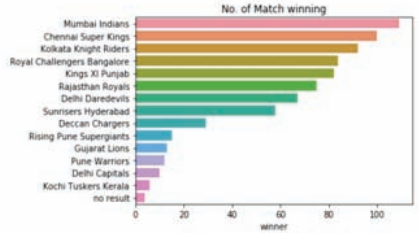
In the Matches Data set, total 756 matches are there and there are total 12 session were played starting from 2008 and the latest session 2019.

The above bubble chart is showing venues with highest session host. Eden gardens hosted 11 sessions, followed by Feroz Shah Kotla, Wankhede, M. Chinnaswamy with 10 session.

Most Successful Team of IPL
In terms of Total match winning

```
#The most successful IPL Team

data = matches.winner.value_counts()
sns.barplot(y = data.index, x = data, orient='h').set(title="No. of Match winning")
```

```
[Text(0.5, 1.0, 'No. of Match winning')]
```



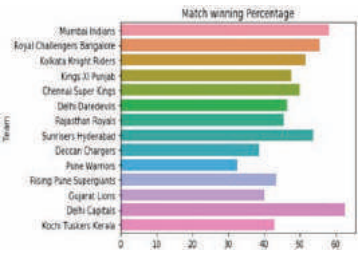

```
Mumbai Indians                109
Chennai Super Kings           100
Kolkata Knight Riders          92
Royal Challengers Bangalore    84
Kings XI Punjab                82
Rajasthan Royals               75
Delhi Daredevils               67
Sunrisers Hyderabad            58
Deccan Chargers                29
Rising Pune Supergiants        15
Gujarat Lions                  13
Pune Warriors                  12
Delhi Capitals                 10
Kochi Tuskers Kerala            6
no result                       4
```

Mumbai Indians is the most successful teams who win 109 matches throughout the 11 sessions (2008-2020). Chennai Super kings is the second most successful team with 100 match win, followed by Kolkata Night Riders with 92 match win. On the other hand Kochi Tuskers Kerala has the least number of match winwith only 6. But one should consider Kochi Tusker Kerala as the most unsuccessful team because Kochi Tusker Kerala played only one session.

In terms of Match winning Percentage:

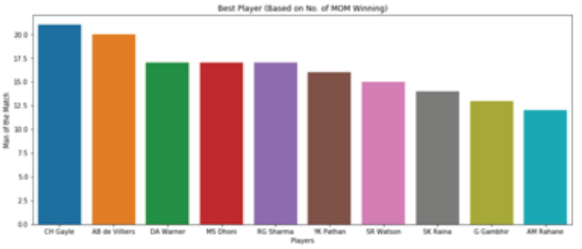
```
matches_total=pd.concat([matches['team1'],matches['team2']])
matches_total=matches_total.value_counts().reset_index()
matches_total.columns=['Team','Total Matches']
matches_total['wins']=matches['winner'].value_counts().reset_index()['winner']
matches_total.set_index('Team',inplace=True)
matches_total.head()

win_percentage = round(matches_total['wins']/matches_total['Total Matches'],3)*100
win_percentage
data = win_percentage
sns.barplot(y = data.index, x = data, orient='h').set(title="Match winning Percentage")
win_percentage
```

```
Team
Mumbai Indians                58.3
Royal Challengers Bangalore   55.6
Kolkata Knight Riders         51.7
Kings XI Punjab               47.7
Chennai Super Kings           50.0
Delhi Daredevils              46.6
Rajasthan Royals              45.6
Sunrisers Hyderabad           53.7
Deccan Chargers               38.7
Pune Warriors                 32.6
Rising Pune Supergiants       43.3
Gujarat Lions                 40.0
Delhi Capitals                62.5
Kochi Tuskers Kerala          42.9
dtype: float64
```



In terms of winning percentage, Delhi Capital is the front runner having 62.5% winning rate, Mumbai Indians stood second with 58.3% ratefollowed by Royal Challengers Bangalore (55.6%).

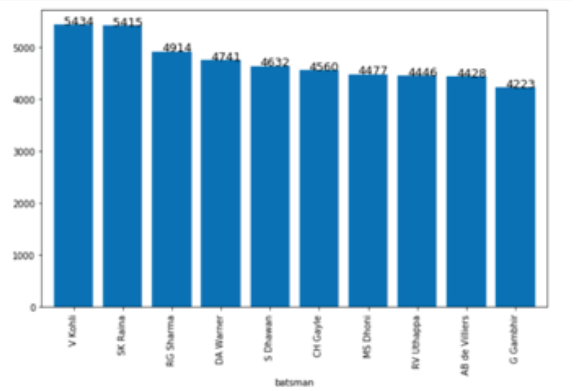
Most Successful Player of IPL
In terms of Man of the Match winner



Above column chart is showing Chris Gayle own highest number of Man of the Match award followed by AB de Villers and David Warner. In case of Indian players, MS Dhoni top the list followed by Rohit Sharma.

In terms of Most Run

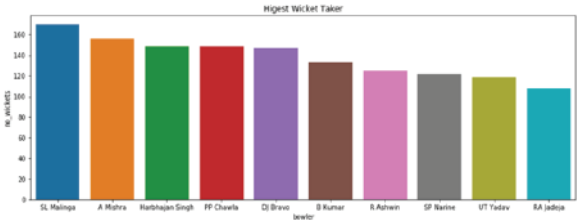
```
#batsman with highest run

plt.subplots(figsize=(10,6))
max_runs=deliveries.groupby(['batsman'])['batsman_runs'].sum()
ax=max_runs.sort_values(ascending=False)[:10].plot.bar(width=0.8)
for p in ax.patches:
    ax.annotate(format(p.get_height()), (p.get_x()+0.2, p.get_height()+2),fontsize=13)
plt.show()
```



In terms of most runs, Virat Kohli is the most successful player with 5434 run across 12 session of IPL. Suresh Raina and Rohit Sharma is in 2nd and 3rd position in this list with 5415 and 4914 runs respectively.

In terms of Most Wickets

```
bowlers = deliveries[["bowler","dismissal_kind"]]
bowlers = bowlers[(bowlers.dismissal_kind.isin(["caught", "bowled", "lbw", "caught and bowled", "stumped"]))]
bowlers = bowlers.groupby("bowler").count().reset_index()
bowlers.rename(columns={"dismissal_kind":"no_wickets"}, inplace=True)
bowlers = bowlers.sort_values(by="no_wickets", ascending=False)
bowlers=bowlers.nlargest(10, ['no_wickets'])
bowlers
plt.figure(figsize=(15,5))
sns.barplot(x=bowlers.bowler, y = bowlers.no_wickets,
            data = matches).set(title="Higest Wicket Taker")
```



In terms of wicket, Lasith Malinga is the best bowler with 170 wickets in his name followed by Amit Mishra and Harbhajan Singh with 156 and 149 wickets respectively.

Most Economical Bowler

```
#Distriabutions of run
eco=deliveries.groupby(['bowler']).sum()
eco['total balls']=deliveries['bowler'].value_counts()
eco['overs']=(eco['total balls']//6)
eco[eco['overs']>200].sort_values(by='overs',ascending=0)['overs'].head(5).reset_index()
eco['economy']=(eco['total_runs']/(eco['overs']))
eco[(eco['overs']>300)].sort_values('economy')[:10].economy.reset_index()
```

	bowler	economy
0	DW Steyn	6.686649
1	R Ashwin	6.754980
2	SP Narine	6.787529
3	Harbhajan Singh	7.043478
4	SL Malinga	7.092929
5	B Kumar	7.237251
6	A Mishra	7.291667
7	PP Ojha	7.404321
8	Z Khan	7.546174
9	P Kumar	7.612756

Economy rate is very important in any T20 games . Economy rate indicates the average runs conceded for each over bowledDale Steyn is the most economical bowler with 6.69 economy followed by R Ashwin (6.75) and Sunil Narine(6.79).

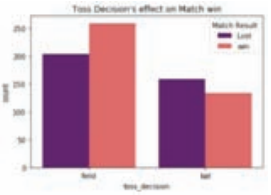
Toss's effect on Match wins:
Toss win vs. Match win:

```
sns.countplot(matches['toss_winner'] == matches['winner']).set(title="Toss win's effect on Match win")
```

From the above Bar chart, it is clear that toss win or loss has least effect to win a match. after winning the toss, team wins the match is higher than when team win the toss but did not manage to win the match.

Toss win vs. Match win

```
#Toss Decision's effect on Match winning
sns.countplot(matches['toss_decision'],hue=(matches['toss_winner'] == matches['winner']),palette = 'magma').set(title='Toss Deci
plt.legend(title='Match Result', loc='best', labels=['Lost', 'Win'])
```



From the above figure, it is clear that Toss Decision's have major impact on Match win.

When team choose field, chance of win matches get increase, naturally choosing batting first increase chance of losing a match.

B. Predictive Analysis

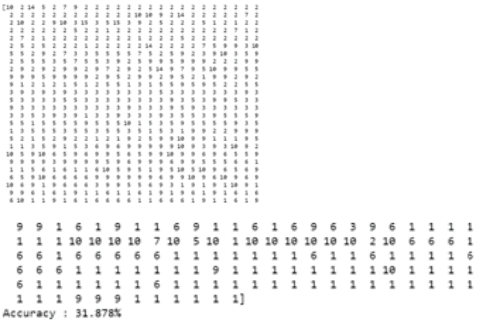
Test Different Machine Learning Model

As this project focus to solve a Classification problem, researchers tested different popular Classification Models to predict outcome of a match based on some dependent variable.

Researchers have used both the team Teams, Venue, Toss Winner, City and Toss Decision as dependent variables.

Logistic Regression Model

```
#Logistic Regression
outcome_var=['winner']
predictor_var = ['team1', 'team2', 'venue', 'toss_winner','city','toss_decision']
model =LogisticRegression()
classification_model(model, df,predictor_var,outcome_var)
```



```
 9  9  1  6  1  9  1  1  6  9  1  1  6  1  6  9  6  3  9  6  1  1  1  1
 1  1  1 10 10 10 10  7 10  5 10  1 10 10 10 10 10 10  2 10  6  6  6  1
 6  6  1  6  6  6  6  6  1  1  1  1  1  1  6  1  1  6  1  1  1  1  6  6
 6  6  6  1  1  1  1  1  1  1  9  1  1  1  1  1  1  1 10  1  1  1  1  1
 6  1  1  1  1  1  1  6  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
 1  1  1  9  9  9  1  1  1  1  1  1]
Accuracy : 31.878%
```

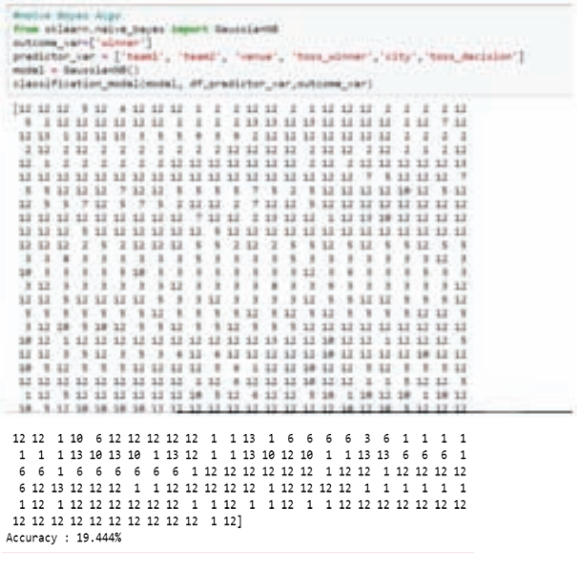
Logistic regression is one of the simplest classification models .the Accuracy of Logistic Regression on this dataset is only 31.878% which is very low and not considerable , that is why researchers should consider other models over logistic Regression model.

K- Nearest Neighbour (KNN) Model

```
#KNN Algorithm
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=3)
classification_model(model, df,predictor_var,outcome_var)
```

K-Nearest is another simple model which based on similarity measure, Distance Function. Applying KNN Model, The Accuracy level increased, now the model has 63.624% Accuracy. The Accuracy level is good but researcher aims to achieve more accuracy.

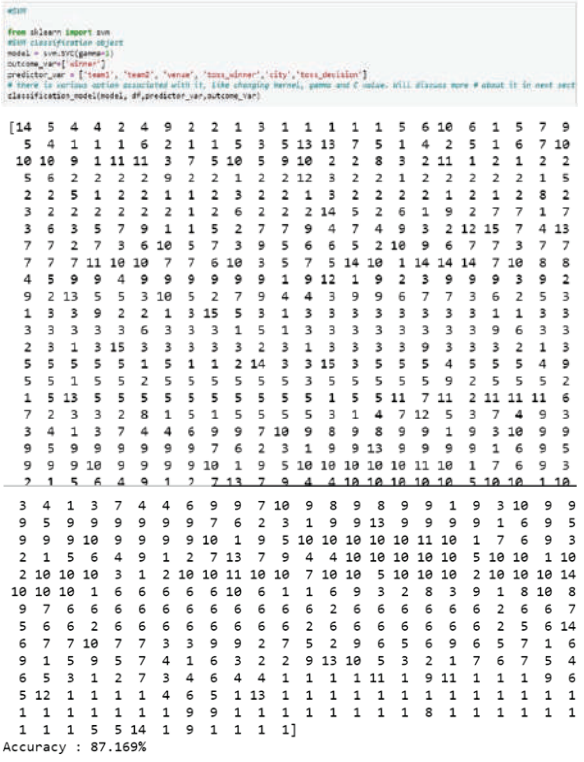
Naïve Bayes Model



Naïve Bayes Model is a classification model, belongs to the probabilistic classifiers group. This model is based on applying Bayes' theorem.

The Accuracy of Naïve Bayes Model is only 19.444%, which is not considerable. Researchers have better option to choose over Naïve Bayes Model.

Support Vector Machine (SVM)



Super vector machine or SVM is used as both classification and regression model.it is a supervised learning model.

The Accuracy of SVM model on the IPL dataset is 87.169% which is very good.

Decision Tree

```
#Decision Tree

from sklearn import tree
model = tree.DecisionTreeClassifier(criterion='gini')
outcome_var=['winner']
predictor_var = ['team1', 'team2', 'venue', 'toss_winner','city','toss_decision']
classification_model(model, df,predictor_var,outcome_var)
```

```
[14  5  4  4  2  4  9  2  2  1  3  1  1  1  1  1  5  6 10  6  1  5  7  9
  5  4  1  1  1  6  2  1  1  5  3  5 13 13  7  5  1  4  2  5  1  6  7 10
 10 10  9  1 11 11  3  7  5 10  5  9 10  2  2  8  3  2 11  1  2  1  2  2
  5  6  2  2  2  2  9  2  2  1  2  2 12  3  2  2  1  2  2  2  2  2  1  5
  2  2  5  1  2  2  1  1  2  3  2  2  1  3  2  2  2  2  1  2  1  2  8  2
  3  2  2  2  2  2  2  2  1  2  6  2  2  2 14  5  2  6  1  9  2  7  7  1  7
  3  6  3  5  7  9  1  1  5  2  7  7  9  4  1  4  9  3  2 12 15  7  4 13
  7  7  2  7  3  6 10  5  7  3  9  5  6  6  5  2 10  9  6  7  7  3  7  7
  7  7  7 11 10 10  7  7  6 10  3  5  1  5 14 10  1 14 14 14  7 10  8  8
  4  5  9  9  4  9  9  9  9  9  9  1  9 12  1  9  2  3  9  9  9  3  9  2
  9  2 13  5  5  3 10  5  2  7  3  4  4  3  9  3  6  7  7  3  6  2  5  3
  1  3  3  9  2  2  1  3 15  5  3  1  3  3  3  3  3  3  3  1  1  3  3
  3  3  3  3  3  6  3  3  3  3  1  5  1  3  3  3  3  3  3  3  3  9  6  3  3
  2  3  1  3 15  3  3  3  3  3  2  3  1  3  3  3  3  3  3  3  3  9  3  3  3  2  1  3
  5  5  5  5  5  1  5  1  1  2 14  3  3 15  3  1  5  5  4  5  5  5  4  9
  5  5  1  5  5  2  5  5  5  5  5  5  3  5  1  5  5  5  9  2  5  5  5  2
  1  5 13  5  5  5  5  5  5  5  5  5  5  1  5  5 11  7 11  2 11 11 11  6
  7  2  3  3  2  8  1  5  1  5  5  5  5  3 12  4  7 12  5  3  7  4  9  3
  3  4  1  3  7  4  4  6  9  9  7 10  9  8  2  8  9  2  1  9  3 10  9  9
  9  5  9  9  2  9  9  9  7  6  2  3  1  9  9 13  9  9  2  3  1  6  9  5
  9  9  9 10  9  9  9  9 10  1  3  5 10 10 10 10 10 11 10  1  7  6  9  3
  2  1  5  6  4  9  1  2  7 13  7  9  4  4 10 10 10 10 10  5 10 10  1 10
  2 10 10 10  3  1  2 10 10 11 10 10  7 10 10  5 10 10 10  2 10 10 10 14
 10 10 10  1  6  6  6  6  6 10  6  1  1  6  9  3  2  8  3  9  1  8 10  8
  9  7  6  6  6  6  3  6  6  6  3  6  2  6  6  6  5  3  6  2  6  6  7
  5  6  1  2  6  1  6  6  6  6  5  6  2  6  6  6  6  3  6  6  2  5  6 14
  6  7  7 10  7  7  3  3  9  9  2  7  5  2  9  6  5  6  9  6  5  7  1  6
  9  1  5  9  5  7  4  1  6  3  2  2  9 13 10  5  3  2  1  7  6  7  5  4
  6  5  3  1  2  7  3  4  6  4  4  1  1  1  1 11  1  9 11  1  1  1  9  6
  5 12  1  1  1  1  4  6  5  1 13  1  1  1  1  1  1  1  1  1  1  1  1  1
  1  1  1  1  1  1  1  9  9  1  1  1  1  1  1 11  8  1  1  1  1  1  1
  1  1  1  5  5 14  1  9  1  1  1  1]
Accuracy : 87.434%
```

Decision tree is a tree like model for easy decision making process.

The Decision Tree Model has almost same accuracy (87.434%) like SVM.

Random Forest

```
#Random forest classifier

model = RandomForestClassifier(n_estimators=50)
outcome_var = ['winner']
predictor_var = ['team1', 'team2', 'venue', 'toss_winner','city','toss_decision']
classification_model(model, df,predictor_var,outcome_var)
```

```
[14  5  4  4  2  4  9  2  2  1  3  1  1  1  1  1  5  6 10  6  1  5  7  9
  5  4  4  1  1  6  2  4  1  5  3  5 13 13  7  5  1  4  2  5  1  6  7 10
 10 10  9  1 11 11  3  7  5 10  5  9 10  2  2  8  3  2 11  1  2  1  2  2
  5  6  2  2  5  2  9  2  2  1  2  2 12  3  2  5  1  2  2  2  2  2  1  5
  2  2  5  1  2  2  1  1  2  3  2  2  1  3  2  2  2  2  1  2  1  2  8  2
  3  2  2  2  2  2  2  2  1  2  6  2  2  2 14  5  2  6  1  9  2  7  8  1  7
  3  6  3  5  7  9  1  1  5  2  7  7  9  4  7  4  9  3  2 12 15  7  4 13
  7  7  2  7  3  6 10  5  7  3  9  5  6  6  5  2 10  9  6  7  7  3  7  7
  7  8  7 11 10 10  7  7  6 10  3  5  7  5 14 10  1 14 14 14  7 10  8  8
  4  5  9  9  4  9  9  9  9  9  9  1  9 12  1  9  2  3  9  9  9  3  9  2
  9  2 13  5  5  3 10  5  2  7  9  4  4  3  9  9  6  7  7  3  6  2  5  3
  1  3  3  9  2  2  1  3 15  5  3  1  3  4  3  3  3  3  4  3  1  1  3  3
  3  3  3  3  3  6  3  3  3  1  5  1  3  3  3  3  3  3  3  3  9  6  3  3
  2  3  1  3 15  3  3  3  3  3  2  3  1  3  3  3  3  9  3  3  3  2  1  3
  5  5  5  5  5  1  5  1  1  2 14  3  3 15  3  1  5  5  4  5  5  5  4  9
  5  5  1  5  7  2  5  5  5  5  7  5  3  5  1  5  5  5  9  2  5  5  5  2
  1  5 13  5  5  5  7  5  5  5  5  5  1  5  5 11  7 11  2 11 11 11  6
  7  2  3  3  2  8  1  5  1  5  5  5  5  3 12  4  7 12  5  3  7  4  9  3
  3  4  1  3  7  4  4  6  9  9  7 10  9  8  2  8  9  2  1  9  3 10  9  9
  9  5  9  9  2  9  9  9  7  6  2  3  1  9  9 13  9  9  2  3  1  6  9  5
  9  9  9 10  9  9  9  9 10  1  9  5 10 10 10 10 10 11 10  1  7  6  9  3
  2  1  5  6  4  9  1  2  7 13  7  9  4  4 10 10 10 10 10  5 10 10  1 10
  2 10 10 10  3  1  2 10 10 11 10 10  7 10 10  5 10 10 10  2 10 10 10 14
 10 10 10  1  6  6  6  6  6 10  6  6  6  6  9  3  2  8  3  9  1  8 10  8
  9  7  6  6  6  6  3  6  6  6  6  6  2  6  6  6  6  3  6  2  6  6  7
  5  6  6  2  6  6  6  6  6  6  6  6  2  6  6  6  6  6  6  6  2  5  6 14
  6  7  7 10  7  7  3  3  9  9  2  7  5  2  9  6  5  6  9  6  5  7  1  6
  9  1  5  9 13 13  4  1  6  3  2  2  9 13 10 13  3  2  1 13  6  7  5  4
  6  5  3  1  2  7  3  4  6  4  4  1  1  1  1 11  3  9 11  1  1  1  9  6
  9  1  5  9 13 13  4  1  6  3  2  2  9 13 10 13  3  2  1 13  6  7  5  4
  6  5  3  1  2  7  3  4  6  4  4  1  1  1  1 11  3  9 11  1  1  1  9  6
  5 12  1  5  1  1  4  6  5  1 13  1  1  1  1  1  1  1  1  1  1  1  1  1
  1  1  1  1  5  1  1  9  9  1  1  1  1  3  1  1 11  8  1  1  5  1  1  1
  1  1  1  5  5 14  5  9  1  1  1  1]
Accuracy : 87.434%
```

Random forest is anensemble type model for classification and regression. In this model, multiple decision trees are constructed to support accurate and quick decision making process.

The Accuracy level of Random Forest is same as Decision Tree which is 87.434%.

Result and Analysis

1. In this it is seen the most IPL matches session took place in Eden Garden and Wankhade stadium which is also a important factor for winning the match.
2. Mumbai Indians is the most successful teams who win 109 matches throughout the 11 sessions (2008-2020).Chennai Super kings is the second most successful team with 100 match wins, followed by Kolkata Night Riders with 92 match win. The Mumbai Indians wins mainly the IPL matches.
3. Mostly the man match of the IPL goes to Chris Gaylewho own highest number followed by AB de Villers and David Warner and in Indian playersMS Dhoni top the list.
4. In this it is clear that toss win or loss has least effect to win a match. After winning the toss, team wins the match is higher than when team win the toss but did not manage to win the match but however the decision to balling or batting first play a major role.

5. In different model of machine learning , we find that Decision Tree, Random Forest and Support Vector Machine have the most accuracy model with around 88% accuracy level.

Conclusion and Scope of Future Studies

From the above studies and analysis researchers Conclude as followings-

Analytics can be used for Cricket match Prediction and it's analysis in very easy way. For IPL game,Teams, Venue, Winning Toss, Venue of the Match and Decision after winning the toss are important influencers to win a match. Different Machine Learning helps to predict outcome of a match. Right selection of Machine Learning Model helps to increase Accuracy of Prediction. From Different Classification Models, Support Vector Machine, Decision Tree and Random forest are best to predict outcome of an IPL games. All of the following gives almost 88% accuracy Level. With this we can predict the IPL match through machine learning models.

Scope of future study

Researchers have taken only few factors as a predictor .but in cricket there are many factors which could have impact on a Match result. Player is one of that influencer. Every year IPL teams use to change their players; lots of new cricketer gets chances to play for the teams. So, in future studies Player can be use as a Predictor. Another important factor is pitch, in future pitch can be use as predictor.

In this study Researchers use only simple popular Classification models, more complex model can be used to increase accuracy in future. Thus using modelling effective decision making can be accomplished through using tools and techniques in sports analytics.

References

[1] *C. Deep Prakash, C.Patvardhan and Sushobhit Singh," A new Category based Deep Performance Index using Machine Learning for ranking IPL Cricketers", Int. Jl. of Electronics, Electrical and Computational System IJEECS ISSN 2348-117X Volume 5, Issue 2 February 2016*

[2] *Parker, David, Phil Burns, and Harish Natarajan. "Player valuations in the indian premier league." Frontier Economics 116 (2008).*

[3] *Singh, Sanjeet. "Measuring the performance of teams in the Indian Premier League." American Journal of Operations Research 1.03 (2011): 180.*

[4] *Saikia, Hemanta, and Dibyojyoti Bhattacharjee. "On classification of all-rounders of the Indian premier league (IPL): a Bayesian approach." Vikalpa36.4 (2011): 25-40.*

[5] *Lenten, Liam JA, Wayne Geerling, and László Kónya. "A hedonic model of player wage determination from the Indian Premier League auction: Further evidence." Sport Management Review 15.1 (2012): 60-71.*

[6] *Rastogi, Siddhartha K., and Satish Y. Deodhar. "Player pricing and valuation of cricketing attributes: exploring the IPL Twenty20 vision." Vikalpa 34.2 (2009): 15-23.*

[7] *Petersen, C., et al. "Analysis of Twenty/20 cricket performance during the 2008 Indian Premier League." International Journal of Performance Analysis in Sport 8.3 (2008): 63-69.*

[8] *http://www.espncricinfo.com/india/content/player/28081.html, T20 statistics of each player*

[9] *http://www.iplt20.com/teams/royal-challengers-bangalore/squad/236/chris-gayle , IPL statistics of eachplayer.*

[10] *http://www.rediff.com/cricket/report/icc-world-cup-de-villiers-maintains-big-lead-shami-rises-to- 7th-in-most-valuable-player-table/20150320.htm*

[11] *C. Deep Prakash, C.Patvardhan and Sushobhit Singh,"A new Machine Learning based Deep Performance Index for Ranking IPL T20 Cricketers", International Journal of Computer Applications (0975– 8887) Volume 137 – No.10, March 2016*

Power System Security Assessment using K-Nnearest Neighbour

P. Praveen*
Dr. C.V.K. Bhanu**

ABSTRACT

Power System Security assessment using conventional technologies is computationally very expensive. Besides, because of large amounts of data to be handled, they can't help operators take rapid decision making preventing major catastrophes like blackouts. Machine learning algorithms with their ability to extract and synthesize information are being increasingly used for applications involved in fast and accurate decision-making including power system security assessment. These algorithms can be used to develop knowledge base from security-constrained optimal power flow algorithms for identifying a security issue in a given power system. The main objective of this work is to demonstrate the application of machine learning technique to assess power system security. The work uses a Test Case for generating data, off-line, and use this data to train k-NN algorithm. After the training, the algorithm predicted the system states with very good accuracy levels. The details of the generated, training, testing and validation are presented in the report. The performance of the algorithm has been validated with different accuracy measures and the same are presented in the results. The result show that the power system accuracy can be predicted using k-NN algorithm. MATPOWER is used in this work for generating off-line data using security-constrained optimal power flow.

Key Words- *k-NN (k-Nearest Neighbour), Security Index, Optimal Power Flow (OPF), MATPOWER*

Introduction

The power system is becoming complex and wide. The power systems main task is to preserve the customers uninterrupted power supply and to retain the power systems working environment in the normal state. To maintain system security is an important factor in the power system operation. System security requires procedures designed to ensure device operation even though the components fail. For instance, due loss of axillary equipment, the generator unit may go down, but b retaining the correct amount of the spinning reserve, the remaining units can cover the deficit and retain the frequency without any shedding. Similarly, a transmission line can fail (damaged b storm and replaced by an automated relay system). If the proper transmission flow is preserved with respect to the generation of commit and dispatch, the persistent transmission lines will handle the increased load and stay within limits [6].

Power system operators endeavor to operate the systems with high degree of security, stability and reliability. Security allude to the capability of the power system to protest for any sudden disruptions without any interruption to customer's service. Which says that there should be sufficient generation as well as transmission resources to meet the load demand projected and also should maintain reserves for any contingencies [3].

As long as the power system was regulated and vertically integrated the systems used were more secure because the grids were designed, built and operated by monopolies and it was also seen that the

*M.Tech Student Scholar, Department of Electrical & Electronics Engineering, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, A.P, India., Email:praveen.padala3@gmail.com
**Professor, Department of Electrical & Electronics Engineering, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, A.P, India, Email: bhanucvk@gvpce.ac.in

generation and transmission are in phase with the increase in loads with limited equipment failures and overloads, which could cause system disturbances. It is also simple to forecast system condition in operations point of view as there are less generation and transmission owners operating in a cooperative manner. However, electric power industry was advanced to open markets over the last decade which incorporate several factors (increased reliance on controls and special protection systems, large numbers of small and distributed generators, etc.) that increase the potential sources for system disturbances, reduces the system robustness and operation predictability. To maintain reliability the system must exactly designed with security as a predominant consideration and monitored for the duration of operation to ensure that there is an adequate margin of protection at all times. The changes that have happened in the new era have changed the need for power system analysis tools and increased the need for more rigorous power system security [3].

The method used to detect whether the system remains in normal (secure) or emergency (insecure) state is referred to as the security assessment of the power system. It is essentially assessing the system's ability to proceed with the provision of service in the event of an unforeseen contingency. The security assessment of the power system is of major concern in planning, design and operation phases of the electrical power system. If the system security is not well defined, the occurrence of certain disturbances may lead to undesirable system emergency conditions and effective control requires rapid safety evacuation. Conventional numerical methods are computationally expensive, making it difficult to use them for online security assessment. Recent developments in machine learning have led to the broad adoption of a number of applications for power systems, ranging from meter data analysis, renewable/load/price forecasting to grid safety (security) assessment. An alternative approach

can be offered by the machine learning techniques for pattern recognition, learning skills and high speed of identifying potential security boundaries.

Background

To determine that the power system is within limits, SCADA systems and state estimator are the primary tools used for measurement. Each transmission system operator's operation system network requires each transmission system operator to classify the operating status of the system. The various operating states of the system are normal state, alarm state, emergency correctable state, emergency state uncorrectable and restorative state [1] as given in Fig. 1.

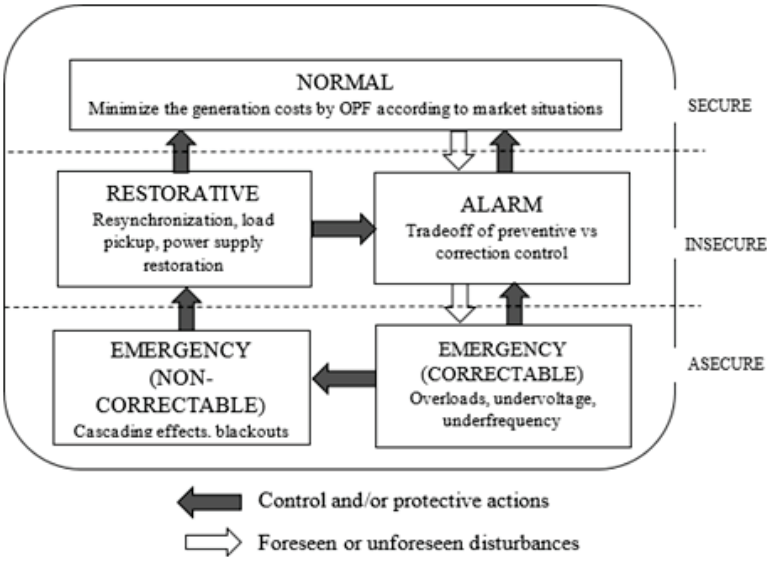


Fig. 1. Operating states and transitions

Usually conventional numerical techniques take time and are therefore not always suitable for real-time applications. Those methods also suffer from the issue of misclassification or/and false alarm. Using machine learning techniques such as k-NN, artificial neural networks, decision trees, deep learning, etc., with their ability to extract and synthesize information, are being increasingly used as an alternative to conventional computational techniques. Security index is calculated with respect the disturbance (N-1 contingency) and this security index helps in finding the state of the system [1].

A. Machine learning

Machine learning allows the machine to learn from data automatically, improve experimental performance, and prophesy things without straightforward programming. A Machine Learning system learns from data, builds models of prediction and predicts the output for it whenever it receives new data. A number of steps such as Problem definition Hypothesis Generation Data Extraction or Collection Data Exploration and Transformation Model Building (Fig2) Model Deployment or Implementation are designed to predict the future model.

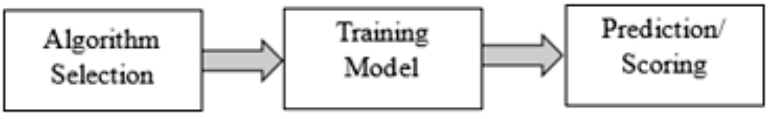


Fig. 2. Steps in model building

A. k-Nearest Neighbor

k-Nearest Neighbor is one among the simplest Machine Learning algorithms which is based on technique of Supervised Learning. k-NN algorithm saves all the available data and classifies new similarity-based data point. This means that when new data appears, it can easily be classified into a suitable category using the k-NN algorithm. k-NN algorithm can be used for both Regression and Classification, but it is mostly used for the Classification problems. It is also called a lazy learner algorithm because it doesn't immediately learn from the training set instead it stores the dataset and performs an action on the dataset at the time of classification [9].

Different performance measures used to measure the performance of the model are accuracy, precision, recall and F_1 score which are obtained from confusion matrix (Fig.3)

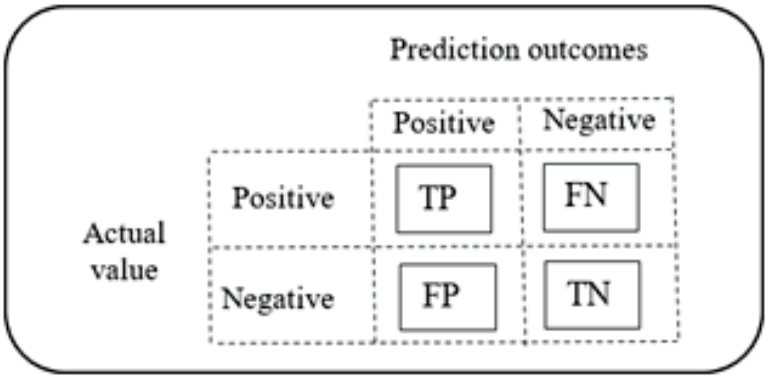


Fig. 3. Confusion matrix

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$F_1 \text{ score} = \frac{2}{\text{Precision}^{-1} + \text{Recall}^{-1}} \quad (4)$$

F_1 score is used when there is confusion whether precision should be given importance or recall is given importance and one should remember that accuracy can be used for a balanced data set. The model is said to be trained accurately if the performance measures are high.

Data Generation

The machine learning algorithms get trained with the help of the available data. The data required for training, testing and validation is obtained from off-line simulation using MATPOWER [9]. Security-constrained optimal power flow is carried out using N-1 contingency criteria on a 5-bus system shown in Fig.4. Security index is calculated and based on its value the system is classified into one of the four states.

To generate data a five-bus system was taken from the MATPOWER. Which consists of five generators, five buses, and six transmission lines with transmission line capacity of 400MW and 240MW for two of the six lines and three loads connected to three different buses. The PJM 5-bus system [10] is shown in the fig. 4.

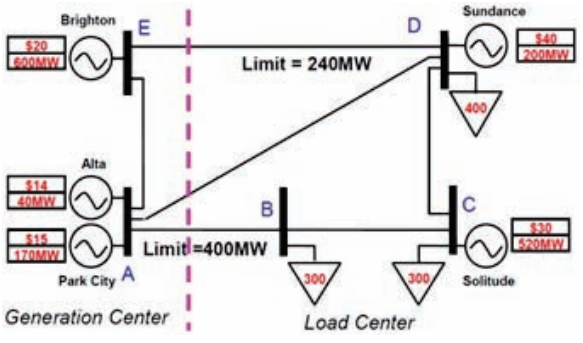


Fig. 4. PJM five bus system

To generate data some changes are done in transmission line capacity of the system, they are considered as AB=400MW, BC=CD=DE=EA= 240MW as an assumption considering there will be no transmission line without having transmission capacity limits. Optimal power flow is run with all the default values. As, MATPOWER Interior Point Solver (MIPS) with 150 maximum number of iterations. In MATPOWER all the cases are having loads which are constant but in real time load is not going to be constant it keeps on changing with respect to the usage. The amount of load varies with time i.e. the load during week days will be different from weekends, load gets changed with respect to seasons and thereby load keeps on changing throughout the year. The change in load can be considered with respect to the IEEE reliability test system-1996 which can be adapted to any system when one desires to model. The IEEE reliability test system-1996 consists of weekly peak load in the

percent of annual peak, the daily load in the percent of weekly peak and the hourly peak load in the percent of daily peak.

Security Index (SI) is the value that says the system security level for a particular operating condition of the system and for a defined contingency. The SI can be calculating with the help of the line overload index (LOI) and voltage deviation index (VDI) [1].

$$LOI_{km} = \begin{cases} \dfrac{S_{km} - S_{lim}}{S_{km}} \cdot 100, & \text{if } S_{km} > S_{lim} \\ 0, & \text{if } S_{km} < S_{lim} \end{cases} \quad (5)$$

$$VDI_k = \begin{cases} \dfrac{|U^{min}| - |U_k|}{|U^{min}|} \cdot 100, & |U_k| < |U^{min}| \\ 0, & |U^{min}| \le |U_k| \le |U^{m}| \\ \dfrac{|U_k| - |U^{m}|}{|U^{m}|} \cdot 100, & |U_k| > |U^{m}| \end{cases} \quad (6)$$

$$SI = \frac{w1 \cdot \sum_{i=1}^{nL} LOI_i + w2 \cdot \sum_{i=1}^{nB} VDI_i}{nL + nB} \quad (7)$$

Where:

- Skm and Slim represents the MVA flows and MVA limits of branch k-m.
- |Umin|, |U | and |Uk| are the minimum voltage limit, maximum voltage limit and bus voltage magnitude of k-bus.
- w1and w2 are the weighing factors of system security.
- nL and nB represents the number of lines and buses.
- The number of data points generated (Fig. 5) for a five-bus system (case 5 in MATPOWER) are 264 (144 branch & 120 generator) points for a day, 1848 (1008, 840) points for a week and a total of 96096 (52416 are by eliminating branches and 43680 are by eliminating generators) data points for a year.

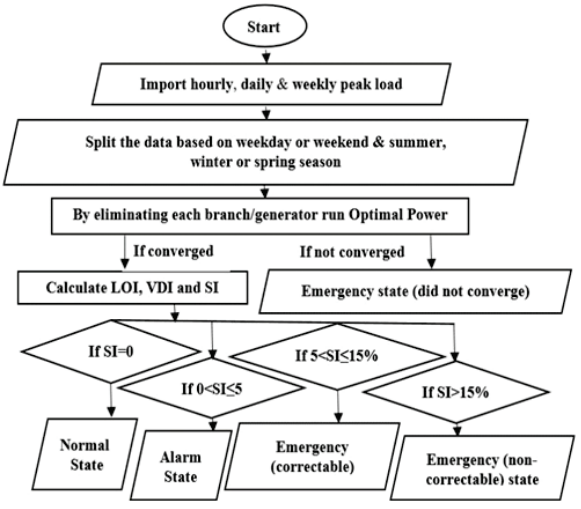


Fig. 5. Basic flow chart to generate data

Implementation and Results

A. Software used

A desktop graphical user interface, Anaconda Navigator included in anaconda distribution allows the user to launch web application Jupyter Notebook where we can use python as a kernel and execute the code

B. Step by Step implementation

- Import all the required libraries.
- load the required data which was generated.
- It should be seen that there are no missing values.
- Segregate the variables to input and target variables.
- Scale all the input variables.
- Divided the total data for training, validation and testing. And seen that all the states are divided equally for training, validation and testing data sets.
- implement k-NN algorithm.

For different values of k (k-neighbors) Mean score and Standard Deviation are calculated by 10-fold cross validation.

While using 10-fold cross validation the data is divided into 10 groups. For each group we get accuracy score. So, we get 10 scores for one k (neighbour) from which we calculate mean score. Mean score is the mean of the 10 scores for one k (neighbour) value. We get a good model from the good score, so the neighbour (k) which is having good accuracy score is selected. Fig 6 represents the mean score for different neighbours.

Fig. 6. Mean score vs. Neighbors

After getting mean validation score standard deviation is calculated. The standard deviation is represented to determine every point deviation corresponding to the mean. The data may not be evenly spread for each and every point. If the data is spread out more than the standard deviation value is higher. Here to train the model in an accurate way it is seen that the standard deviation is minimum for the corresponding neighbour with good mean validation score. Fig 7 shows the standard deviation for different neighbours.

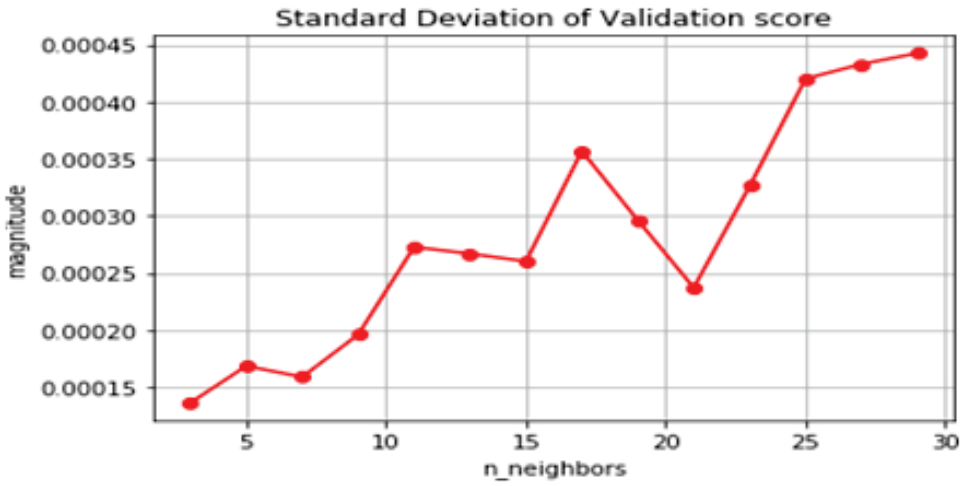


Fig. 7. Standard deviation vs. Neighbors

F_1 score is calculated for Train Data set and Validation Data set, for different values of k (neighbours) which is represented in Fig 8.

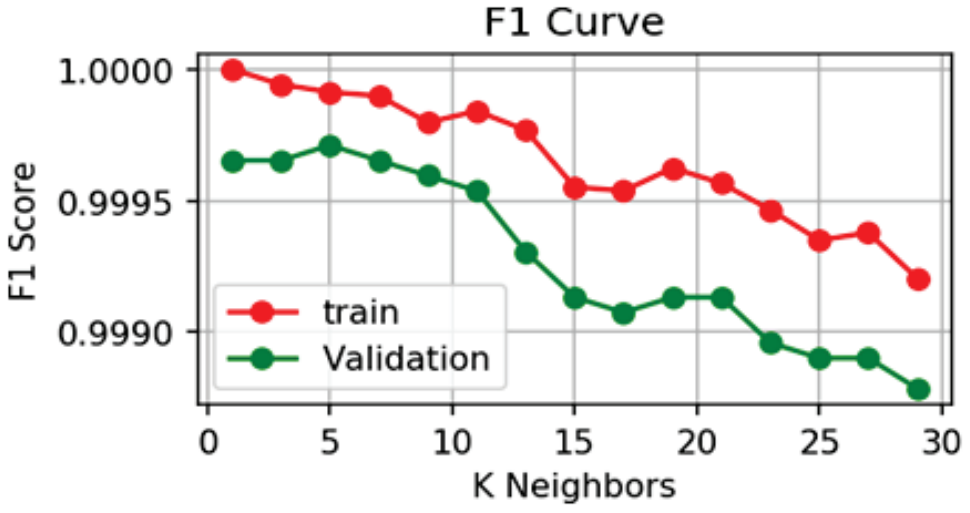


Fig. 8. F_1 Score vs. Neighbors

Under-fit, over-fit and best-fit can be determined by considering the graph of F_1 Curve which is drawn for train and validation data set. If the validation score is moving away from the train score it is considered as over-fit (Memorizing the lessons). If the validation score is running around parallel with the train score it is considered as under-fit (not interested in learning). If the validation score is moving towards the train score it is considered as best-fit (conceptual learning).

To determine the best neighbor (k) to get an accurately predicted output maximum mean score (Fig.6), minimum standard deviation (Fig.7) and best-fit (Fig.8). Here by considering the mentioned three k value (neighbor) is taken as 5. By taking k value as 5 the F1score obtained for validation data set is 0.9997107374995196.

A. Results

For a test case of 5 bus system, considering N-1 contingencies for one year i.e. 8,736 hours, a total of 96,096 data points was generated. The generated data consists of 25076 samples of normal state, 54283 samples of alarm state, 16671 samples of emergency correctable state and 66 samples of emergency states which did not converge. This generated data was treated for missing values and was divided into datasets and used in k-NN algorithm. Depending upon the training data set (69141) and validation data set (17286) k value is taken as 5 and applied for testing. To get the desired output we need to input seven values consisting of five bus voltages, total generation and total load (as shown below); which is not an easy task. So, a test dataset

consisting of 9603 data samples which was not used anywhere in the k-NN model development was used as an input.

When test data was applied as an unknown new input, the following results were obtained. The confusion matrix (Table I) obtained consists of actual values and predicted outcomes of the three states of the system i.e. Normal state, Alarm state and Emergency correctable state.

TABLE I

TEST data	Prediction outcomes		
Actual values	5425	0	0
	1	1667	0
	2	0	2508

The confusion matrix helps in drawing different conclusions about the number of samples which are true positive, true negative, false positive and false negative. From confusion matrix it is every clear that how many samples are predicted correctly and how many samples are not predicted correctly. The details of correctly predicted and wrongly predicted samples are given in the Table II.

TABLE II

STATE of the system	No.of Samples	Correctly predicted	Wrongly predicted
Normal	2510	2508	2
Alarm	5425	5425	0
Emergency controllable	1668	1667	1

The precision, recall and F1 scores of the normal state, alarm state and emergency state samples as an unknown input are shown in Table III.

TABLE III

State	Precision	Recall	F_1 -score	Support
Alarm	0.99944731	1.00000000	0.99972358	5425
Normal	1.00000000	0.99940048	0.99970015	1668
Emergency correctable	1.00000000	0.99920319	0.99960143	2510

The overall model accuracy, precision, recall and F1-score for the specified neighbor (k=5) in k-NN algorithm.

Accuracy Score = 0.999687597625742 Precision Score = 0.9996877702873341 Recall Score = 0.999687597625742 F_1 Score = 0.9996875836865246

Conclusion

In this thesis k-NN algorithm is implemented for Power System Security Assessment on a 5-bus system. N-1 contingency criteria is used for arriving at the severity index explained in the methodology. Based on the security index, the system has been classified into four states – normal state, alert state, emergency (correctable) state and emergency (non-correctable) state. The number of cases are 96030 (from the generated 96096, as the present work does not consider emergency – non-correctable state). From the total sample of 96030 with 72% training data, 18% validation data and 10% testing data. The performance of the k-NN algorithm used is indicated with different metrics which measure the accuracy of the system for the test data (accuracy score=0.999, precision score=0.999, recall score=0.999 and F_1 score=0.999) which convey that the model is well trained and can be implemented for security of the system.

References

[1] *Tomin NV, Kurbatsky VG, Sidorov DN, Zhukov AV. "Machine learning techniques for power system security assessment". IFAC-Papers on Line 49(27) pp. 445–50, 2016.*

[2] *V. Kurbatsky, N. Tomin, "Identification of pre-emergency states in the electric power system on the basis of machine learning technologies", 2016 12th World Congress on Intelligent Control and Automation (WCICA), pp. 378-383, 2016.*

[3] *Wang L, Morision K, Kundur P "Power system security assessment". IEEE Power & Energy Magazine 2004:2(5):30-39.*

[4] *Oliver Theobald, "machine learning for absolute beginners". Second edition.*

[5] *Raul Garreta, Guillermo Moncecchi "Learning scikit-learn: Machine Learning in Python". Packet publishing open source community experience distilled.*

[6] *Allen J.Wood, Bruce F Wolenberg, "Power Generation, Operation and Control". John Wiley and Sons, Inc.*

[7] *U.G.Knight, "Power Systems in Emergencies". John Wiley and Sons, ltd.*

[8] *Online Machine Learning course by INTERNSALA TRAININGS. https://trainings.internshala.com/progress/home/machine-learning.*

[9] *Ray D. Zimmerson, Carlos E. Murilo-Sanchez, "MATPOWER User's Manual Version 7.0". June 20, 2019.*

[10] *F.Li and R.Bo, "Small Test Systems for Power System Economic Studies", Proceedings of the 2010 IEEE Power & Energy Society General Meeting.*

Sementic Information Retrieval: Unboxing the Complexity in Big Data

Khushbu A Patel*
Dr. Mahammad Idrish I. Sandhi**

ABSTRACT

Big data has now become a new normal in the data analytics field. Many institutions and organisations including data analytics companies, governments organisations, financial institutions, health care providers etc. have started resorting to big data techniques for analysis, business enhancing strategies and providing better customer experience. Big data are a large sets of complex, complicated big volume data that help in analysing, discovering and processing uncovered, disorganised, complex and valuable information that can be useful in enhancing business prospects and growth strategies. Big data includes three Vs i.e. Velocity, Variety and Volume. This excel the data analysis capacity and techniques bigger than the traditional relational database methods. For this, it requires schema-less techniques for processing and analysing such big data. Without adjoining semantic information retrieval techniques, it cannot become the great solution. This calls for extracting information through semantic information retrieval techniques instead of using structural information only. This calls for semantic relevancy. The semantic information retrieval systems have to adapt to as per the domain knowledge. While using semantic information retrieval techniques, queries should be constructed that could be useful in fetching semantically useful information or documents as compared to syntactic retrieval. Interlinking of information or documents happens through semantic information retrieval among various documents or sources so that a holistic view of domain could arise. Knowledge graph could be used to arrive at the solution of interlinking problems that have been obtained from the semantic interlinking. Efforts have been made to inculcate and analyse various semantic information retrieval techniques with their which are useful in big data.

Introduction

Big data simply mean data which are in large amount, difficult to process using traditional processing techniques, which are complex in nature and require large amount of storage for storing large amount of information. What big data comprises is not much important, it is more important that what companies do with such large amount of data, i.e. big data. Many stock exchanges, social media site companies any many more generate more than 500 terabytes of data per day. These data, if combined, become petabytes in amount and are difficult and complex to process using traditional processing technologies. Big data which may be in the form of structured, unstructured or semi-structures form, requires big and immersive infrastructure which is costly to handle these amount of data. Schema-less architectures may be useful in processing these data which may give uncompromised quality in data processing. Machine learning models may use metadata which are formed using any raw data. By looking at the various patterns of data at different intervals, an Artificial Intelligence models could be useful to process these data. (1) For processing big data, automation of existing information techniques is required but it is very difficult, costly and it requires immersive efforts on the part of human. Information Retrieval (IR) system, which is a part of communication system, focuses on knowledge filtering could be useful in processing big data. The main pillar of this system is to disseminate

*Assistant Professor, Shree Uttar Gujarat BCA College, Surat, Contact: khushu.bca@gmail.com
**Associate Professor & Head , Sankalchand Patel College of Engineering Department of Computer Application, Sankalchand Patel University, Visnagar, Contact: idrish.mca@gmail.com

information to the right user, at the right time and right information what the user exactly needs. IR system disseminate this information based on the domain knowledge. IR systems may be evaluated based on their efficiency of transforming query to an opportunity of search tasks. The retrieved document should be ultimate end result of the user and it should be exactly as per the user requirements. It should be disseminated within the required time frame. If IR system disseminates the user requirements within the time frame, it is said that this system is efficient and vice versa. Thus, end user is the sole element of information retrieval system. (2) Semantic information system can be useful in retrieving domain knowledge. It creates knowledge graph which is composed of connecting similar data points. This indicates the whole picture of the domain knowledge. If domain knowledge becomes difficult because of information overload, sematic information retrieval technique is useful in such a situation. Semantic information system uses query, machine learning and other techniques for retrieving information. It uses caches up the query entered in the system and goes through it with the information context. In machine learning, the machine understands the user's requirements from his point of views and disseminate that information which is required

by him. User usually prefers this type of system as it provides information based on his preferences. This paper tries to study various IR approaches which have been useful in disseminating information with semantic information retrieval system. Focus is also given on various information filtering techniques, advancements in information retrieval and benefits and threats to these systems.

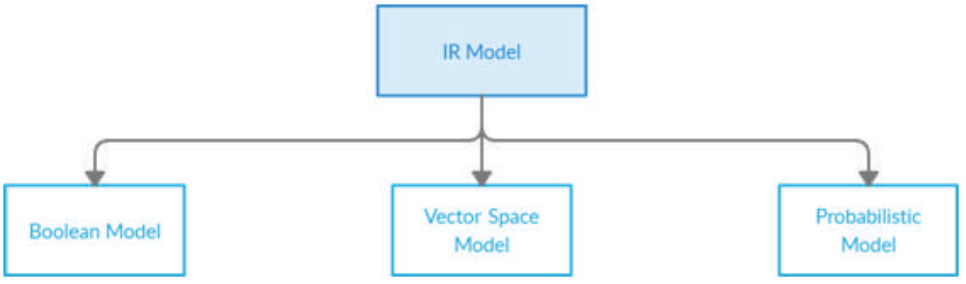


Fig 1: IR Model

Literature Review

Kai-Mo Hu, Bin Wang,Jun-Hai Yong,Jean-Claude Paul (2013) studied that the performance of retrieval techniques can be enhanced with VSM through Latent Semantic Indexing. Keyword based similarity techniques can be made effective by LSM as it can be adapted to synonymy with ease. By aggregating semantic analysis with raw VSM, the performance of SIR techniques can be improved. Priori collapsing facilitates avoiding words of different parts of speech in the syntactic analysis. Commercial establishments uses CAD (Computer Aided Design). This leads to massive increase in database assemblies. VSM is used to improve the assemblies of these database. (3).

Fouad Dahak, Mohand Boughanem, Amar Balla, (2016) while studying "a probabilistic model to exploit user expectations in XML information retrieval" concluded that for document ranking in IR system various techniques are used like logical model, language model, semantic indexing techniques etc. They found that the performance of retrieval function could be improved through term proximity information. It has the structure of tree that can be used for XML document retrieval. (4)

Daniel Z. & Zanger, (2002) have studied IR models which are categorised as Boolean Model, Vector Space Model and Probabilistic Model. They derived IR as to filter document information from the domain in lieu of user query. They identified that many corporate users

use Boolean Model. They also identified some weaknesses in these models like errors of efficiency, query subjectivity etc. which are common issues in SIR techniques. (5)

Daifeng Li, Andrew Madden, (2019) found that because of the complex nature of retrieval data types, case based reasoning becomes more difficult that obstructs the similarity computation. Thus, Bayesian method focuses on probabilistic inferences instead of similarity calculations. They also concluded that in order to achieve parameter independence test, the Bayesian Network technique is used with WC algorithm to assign computational task in big data. (6)

Nigel G. Ward, Steven D. Werner, Fernando Garcia, Emilio Sanchis, (2015) also studied VSM model to improve assembly retrieval. In this, based on query similarity ranking of documents happens on the basis of cosine ranking formula. They also studied that in order to match bipartite graph matching problems, assembly retrieval technique can be extended with part matching algorithm. (7)

A.G. Lopez-Herrera, E. Herrera-Viedma, F. Herrera (2009) while studying Applying multi- objective evolutionary algorithms to the automatic learning of extended Boolean queries in fuzzy ordinal linguistic information retrieval systems, found out IR systems as the very emerging techniques. Boolean operators and relevance ranking methods are used in natural language processing. In batch retrieval which is called evaluation technique, various user documents, queries are used to analyse the efficiency of various techniques. The relevance score can be optimised through query expansion. (8)

Ben He, Jimmy Xiangji Huang, Xiaofeng Zhou (2011) proposed probabilistic IR model and estimation technique. They researched that for efficient retrieval and exact tracking, two components namely probabilistic IR model and estimation technique are responsible. Relevance feedback is facilitated by probabilistic IR model with DCM as it uses negative feedback to improve retrieval algorithm. (9)

Karen e. Lochbaum and Lynn A. Streeter (1989) while comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval, have identified that documents and user queries represent vector as techniques. They have also analysed that when different word documents have different meaning, word based similarity poses errors. (10)

Zuobing Xu, Ram Akella (2010) studied that term dependency based retrieval model can be used to improve linked dependence assumption. In this, the query can be expanded through dependency structured index system and chow expansion. To derive the dependency in the query terms, dependency parser can be used. IR system may lead to the problems of estimation of relevant and irrelevant classes. (11)

Mourad Sarrouti, Said Ouatik El Alaoui, (2017) have also found that in order to minimise errors of relevance and granularity, query dependent and independent features are applied. (12)

Semantic Information Retrieval Techniques

This fig. 2 explains that the Semantic Information Retrieval techniques can be categorised into 2 heads, i.e. Indexing Techniques and ontology Search Techniques. These techniques are explained in this paper in the details. Semantic analysis is to ensure that the statements of the programs should be semantically correct. Syntax tree and symbol table are used in Semantic Analysis to ensure that the program is semantically consistent with basic input of language definitions. The main problem in natural language processing is the semantic analysis. Words will have different meaning in different context as they have syntactic relationship with one another. It creates an ambiguity in the processing of requests. This calls for semantic information retrieval techniques. SIR techniques should be able to use the valid extraction which is relevant to natural language information which lead to better diagnose language query. (13)

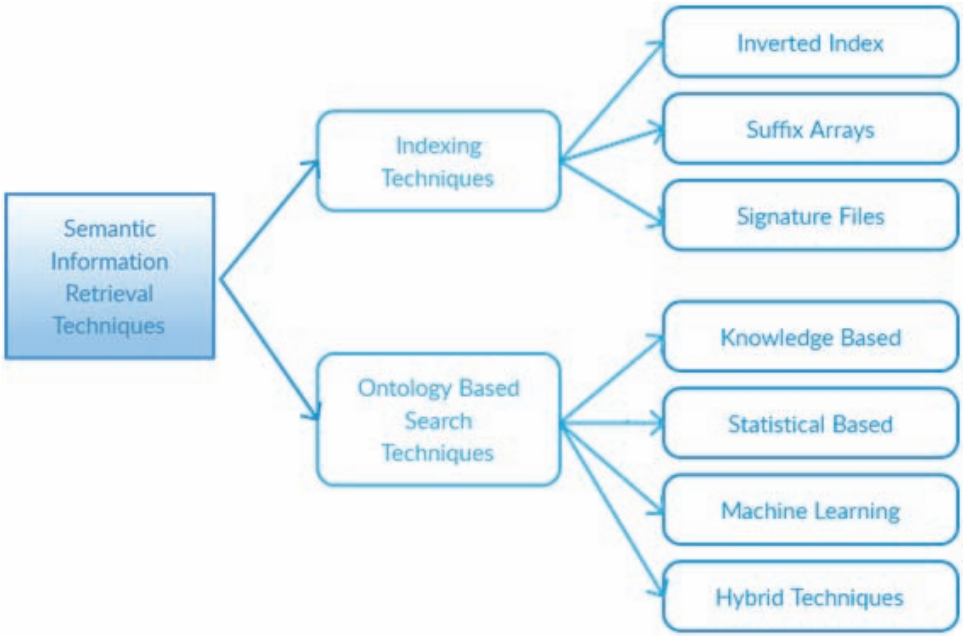


Fig 2: Semantic Information Retrieval Techniques

INDEXING TECHNIQUES:

When the results or queries become more complex, semantic analysis will first reduce the similarity between the query elements. For this process, the best indexing techniques are called for which let users to model entities and keywords. In order to customise the data structure, this type of indexing techniques are required to adjust Boolean retrieval and ranking methodology. Semantic information with keywords and queries can be retrieved using Boolean retrieval techniques. It requires a large amount of data storage and time on the part of users. Thus, for assessing the quality of this type of techniques, data storage, speed of the construction and time taken are taken into the consideration. (14) Search engines major like Google, and other ecommerce companies like amazon use indexing techniques. It matches the keyword entered in the query with the words in its database. It retrieves the matching document or result and ranks it using a ranking algorithm. (15) SemIndex+ is the semantic indexing scheme for constructing structured, unstructured and semi-structured data.

Inverted Index:

Extension of the inverted file structure is used to construct related inverted files. Semantically related words in the inverted index can be appended to build inverted files structure. Page rank values are used to determine semantic relation which can lead to the improvement in the system efficiency. (16) Inverted index represents every word in the document as a keyword. The recording of every word is done with the respective document which leads to a corresponding location. Thus, a query containing the specific keyword is presented in the inverted index, the translation is done into the specific location (17). When a document is added in to the database, inverted index allows the user to do fast full text searches but at the same time, the cost of processing will increase. The number of pages related to a keyword change and the web document content also gets changed when inverted index has been put into the dynamic space. By forming the semantic knowledge base and constructing hybrid retrieval model, inverted index can be expanded further. SemIndex is this type of hybrid inverted index (18).

Suffix Array:

Computing terms and document frequencies can be eliminated through suffix trees. The longer the n-grams, higher is the precision score of IR system. It

has been researched that in the coming future, by lengthening weighting, can improve IR performance with longer n-grams (19). Suffix trees have been defined as the tree based data which can be used to solve string related problems (20). Suffix tree and suffix array both are similar terminologies but they can

be differentiated in terms of space they occupy. Suffix array occupies very less space than suffix tree. String's starting point is called suffix and in this, searching is done through binary search. In suffix tree, there is a trade-off between word segmentation and information retrieval performance but, it doesn't guarantee efficient performance. Suffix arrays come in to the rescue. Compressed data structure for pattern matching is called Compressed Suffix Array. This could be a topic of research in the field of Semantic Information Retrieval.

Signature Files:

The compressed version of database which is created as an abstraction of a document and collection of these documents is called Signature files (21). Bit strings, included in the document are used to construct signature files, which are used by large databases for making efficient information filtering. In making a query, only signature files are searched and other files are automatically rejected by the process. But, performance failures often occur when signature weights are not up to the mark. To

Algorithm 4 Tabu Search

```
1: Empty the tabu list
2: Randonly create the initial solution s
3: While the termination criterion is not met
4:     v = NonTabu-NeighborSelection(s)
5:     If v satisfies the improving conditions
6:     s = v
7:     Update the tabu list based on s
8: End
```

avoid this type of failure, frame slice technique should be applied to create the hierarchical signature files. This type of approach helps making retrieval of information more efficient (22). Multilevel hashing and frame slicing techniques are often used to construct two dimensional dynamic signature file in order to achieve dynamic storage (23).

Ontology Based Search Techniques

There has been an increase in the online learning resources through learning management systems recently. Ontology is also used in recommender programmes. For the presentation of a knowledge of a domain or part of it, ontology is used in various fields like semantic web, AI, systems engineering, information architecture, library science etc. ontology is the core system in any domain for representing information of that particular domain, and in the absence of ontology, vocabulary of knowledge representation cannot exit. Very large amount of data has been created in the form of n number of links in a given query through the web page interlinking which also produce most relevant search result for the users (24).

Knowledge based techniques:

In order to retrieve exact response to user queries in question answering system, the user questions are converted to multiple levels of transformation. Thus, in an Information Retrieval system, exact responses are developed instead of the list of multiple responses. For this type of problems in RDF databases, triple extraction algorithm is efficient (25). While designing a product, semantic retrieval in mechanical domain is crucial (26). Because of various problems inherent in this domain, like vast domain knowledge, traditional keyword search or semantic techniques are not sufficient for information retrieval. In this field, ontology comes to rescue. Ontology itself includes the whole knowledge domain, thereby it increase the efficiency of information retrieval. The simple keyword query is converted into the Boolean Model, and weights are assigned to each entered keywords. The domain ontology helps in making traditional semantic keys into sophisticated semantic keys which are more

efficient in retrieval. But, the more the extension of query, the more are the errors in retrieval (27). The data model must be semantic while constructing ontology for big data. Domain knowledge and other related entities should be included in the ontology. To enhance the efficiency of domain ontology, metadata could be entered in to the knowledgebase of ontology. However, negative side of ontology is but not limited to mapping between different ontology, loss of data while tr4ansformatoin process of ontology to database etc.

Statistical Techniques:

In the statistical ontology, only statistical information included in the database can be extracted with the statistical techniques. Statistical techniques of ontology are probabilistic techniques and can be applied after the pre-processing stage. Concept extraction and taxonomical extraction are the key areas where statistical techniques of ontology work. Ontology is used in a large number of situations but the one weakness in using it is learning automation. Statistical techniques of ontology are used to derive the meaning of domain, conce4pts and other relations of these terms (28). In ontology learning, natural language processing is used in all the stages of processing.in order to use ontology in a formal manner, Inductive Logic Programming is used. It derives the meaning of domain and makes other algorithms easy for the presentation. The precision value is increased when the C value technique is used in the ontology technique. Real terms used to top the list and helps in processing multi-word documents through the usage of contextual information (29). For achieving auto learning, two approaches are assumed to be appropriate and they are Singular Value Decomposition and Latent Drichlet Allocation. To build topic ontology, these methods extracts statistical relationship among various terms and documents. Based on this relationship, ontology graph is generated. To provide effective knowledge base, terminology ontology is constructed for optimizing semantic query.

Machine Learning Techniques:

Ontology web language and rule-based sophisticated language are most important now a days. Various terminological sources are used to include semantic information through multi-level classifiers. Through the measures like KL-distance and others, the effectiveness of ML based IR systems can be improved. Classification algorithms in machine learning which are used to split data into classes are also used to figure out to mapping weight between different entities in ontology (30). Three stages are required in semantic search, i.e. selection of apt resources, query alteration and ranking retrieved data. Machine learning classifiers and terminology based classifiers are used in ontology mapping.

Hybrid Techniques:

Semantic closeness measure is used is to improve the effectiveness of retrieval technique (31). Fuzzy logic is useful in identifying query uncertainty through constructing hybrid ontology. Formal Concept Analysis (FCA) is useful in auto-creating fuzzy logic ontology through clustering. Semantic value concepts in ontology or knowledge graphs are ignored in hybrid techniques. In this situation, the derived inferences are not always true, they may provide wrong inferences. Based on dynamic performance analysis, relevance performance algorithms can be used to update the system. Thus, after the testing, the whole system becomes learned the appropriate configurations and preferred paths for a particular domain (32). The query translation becomes difficult to run when the user query is complex in question answering system.

Conclusion

Instead of syntactic retrieval semantically retrieved documents could be fetched by constructing queries in semantic information retrieval. The efficiency of IR systems can be reflected in its efficiency of transformation of queries into the meaningful search results. Relevance score can be tackled through query expansion. For query expansion and result ranking, knowledge graphs are most important for IR systems. Accurate query execution can be done through Ontology based Semantic Retrieval Techniques. Statistical ontology techniques have been used for effective optimisation of search techniques. Hybrid techniques of ontology is also used immensely for search optimisation. Thus, any individual wanting semantic search techniques can be advised through this paper as it helps in selecting SIR techniques which are suitable for individual needs.

References

1) *R. Priyadarshini, Latha Tamilselvan, T. Khuthbudin, S.Saravanan and S. Satish, (2015)" Semantic Retrieval of Relevant Sources for Large Scale Virtual Documents", Procedia Computer Science 54, 371 –379.*
2) *Antonio M. Rinaldi, Cristiano Russo, (2018)" User centered Information Retrieval using Semantic Multimedia Big Data", DOI:10.1109/BigData. 2018.8622613.*
3) *Kai-Mo Hu, Bin Wang,Jun-Hai Yong,Jean-Claude Paul, (2013)" Relaxed lightweight assembly retrieval using vector space model", Computer-Aided Design 45 739– 750.Elseiver*
4) *Fouad Dahak, Mohand Boughanem, Amar Balla, (2016)" A probabilistic model to exploit user expectations in XML information retrieval", Information Processing and Management 1–19.*
5) *Daniel Z. & Zanger, (2002)" Interpolation of the extended Boolean retrieval model", Information Processing and Management 38 743–748.*
6) *Daifeng Li, Andrew Madden, (2019)" Cascade embedding model for knowledge graph inference and retrieval", Information Processing and Management 56 102093, Elseiver.*
7) *Nigel G. Ward, Steven D. Werner, Fernando Garcia, Emilio Sanchis, (2015)" A prosody-based vector-space model of dialog activity for information retrieval", Speech Communication 68 85–96.Elseiver.*
8) *A.G. López-Herrera, E. Herrera-Viedma, F. Herrera, (2009)" Applying multi-objective evolutionary algorithms to the automatic learning of extended Boolean queries in fuzzy ordinal linguistic information retrieval systems", Fuzzy Sets and Systems 160 ,2192– 2205.*
9) *Ben He, Jimmy Xiangji Huang, Xiaofeng Zhou, (2011) "Modeling term proximity for probabilistic information retrieval models", Information Sciences 181 3017–3031.*
10) *Karen e. Lochbaum and Lynn a. Streeter,(1989)" comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval" Information Processing & Management Vol. 25, No. 6, pp. 665-676.*
11) *Zuobing Xu, Ram Akella, (2010)" Improving probabilistic information retrieval by modelling burstiness of words", Information Processing and Management 46 143–158. Elseiver.*
12) *Mourad Sarrouti, Said Ouatik El Alaoui, (2017) "A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering", Journal of Biomedical Informatics 68 96–103. Elseiver.*
13) *Mukundan Karthik, Mariappan Marikkannan, and Arputharaj Kannan, (2008) "An Intelligent System for Semantic Information Retrieval Information from Textual Web Documents", IWCF, LNCS 5158, pp. 135–146. Springer.*
14) *J. Tekli, R. Chbeir, A.J.M. Traina (2018), "SemIndex+: A semantic indexing scheme for structured, unstructured, and partly structured data", Knowledge Based Systems https://doi.org/10.1016/j.knosys.2018.11.010. Elseiver.*
15) *Fatemeh Lashkari, Faezeh Ensan, Ebrahim Bagheri, Ali A. Ghorbani, (2016), "Efficient Indexing for Semantic Search", Expert Systems With Applications, doi:10.1016/j.eswa.2016.12.033. Elseiver*
16) *Richard Chbeir, Yi Luo, Joe Tekli,_, Kokou Yetongnon, Carlos Raymundo Ibanez, Agma J. M. Traina, Caetano Traina Jr.,and Marc Al Assad,(2014)" SemIndex: Semantic-Aware Inverted Index", ADBIS, LNCS 8716, pp. 290–307, Springer International Publishing Switzerland.*
17) *Shaojun Zhong, Min Shang and Shijuan Deng, (2011) "Design of the Inverted Index based on Web Document Comprehending", Journal of Computers, vol. 6, no. 4, Elseiver.*
18) *Ben Carterette, Fazli Can, (2005) "Comparing inverted files and signature files for searching a large lexicon", Information Processing and Management 41, 613–633, Elseiver.*
19) *Nieves R. Brisaboa, Ana Cerdeira-Pena, Antonio Farina, and Gonzalo Navarro, (2015)" A Compact RDF Store Using Suffix Arrays", SPIRE, LNCS 9309, pp. 103–115, 2015. DOI: 10.1007/978-3-319-23826-511.*
20) *Jin Hu Huang and David Powers, (2008) "Suffix Tree Based Approach for Chinese Information Retrieval", Eighth International Conference on Intelligent Systems Design and Applications, DOI 10.1109/ISDA.2008.365,IEEE.*
21) *Jeong-ki kim Choon-hee lee, Jae-Woo Chang, (1997). "Two-Dimensional Dynamic Signature File Method Using Extendible Hashing and Frame-Slicing Techniques", INFORMATION SCIENCES 98, 1-26, Elseiver.*

22) *Jeong-Ki Kim, Jae-Woo Chang, (2000) "Vertically partitioned parallel signature file method", Journal of Systems Architecture 46, 655-673. Elseiver*

23) *Byoung-Mo IM, Myoung Ho Kim, Jae Soo Yoo, Kil Seong Choi, (1999) "Dynamic Construction of Signature ®les based on frame sliced approach", Data & Knowledge Engineering 30 101-120. Elseiver.*

24) *Prerna Parmeshwaran, Juilee Rege, Sindhu Nair, (2015) "The Use of Ontology in Semantic Search Techniques", International Journal of Computer Applications (0975 – 8887) Volume 127 –No.6.Elseiver.*

25) *Kamran Munir, M. Sheraz Anjum, (2018) "The use of ontologies for effective knowledge modelling and information retrieval", Applied Computing and Informatics 14,116–126.*

26) *Songhua Ma and Ling Tian, (2015) "Ontology-based semantic retrieval for mechanical design knowledge", International Journal of Computer Integrated Manufacturing, Vol. 28, No. 2, 226–238, Elseiver.*

27) *Amol N. Jamgade and Shivkumar J. Karale, (2015) "Ontology Based Information Retrieval System for Academic Library", 2nd International Conference on Innovations in Information, Embedded and Communication systems . IEEE.*

28) *Asim, M.-N., Wasim, M., Khan,M.U.G. (2018), "A survey of ontology learning techniques and applications", Database, doi:10.1093/database/bay101.*

29) *Monika Rani, Amit Kumar Dhar, O.P. Vyas, (2017), "Semi-automatic Terminology Ontology Learning based on Topic Modeling", Engineering Applications of Artificial Intelligence, Volume 63, Pages 108125, Elseiver.*

30) *Umberto Straccia and Raphael Troncy, (2006) "Towards Distributed Information Retrieval in the Semantic Web: Query Reformulation Using the Omap Framework", ESWC, LNCS 4011, pp. 378–392. Springer.*

31) *Balasubramaniam K, (2015) "Hybrid Fuzzy-Ontology Design using FCA based Clustering for Information Retrieval in Semantic Web", 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50 135 – 142, Elsevier*

32) *Monika Rani, Maybin K. Muyeba, O.P. Vyas, (2014), "A hybrid approach using ontology similarity and fuzzy logic for semantic question answering." In Advanced Computing, Networking and Informatics-Volume 1, pp. 601-609. Springer.*

Social Media Analytics Platform to assist Business Decision Making for Small and Medium Enterprise in Indonesia

Muhammad Apriandito Arya Saputra*
Manahan Siallagan*
Santi Novani*
Lidia Mayangsari*
Yogie Setiafriawan*
Puteri Annisa Tsamrotul Fuadah*

ABSTRACT

In today's big data era, many big companies are competing in utilizing social media data to support the business strategy decision making processes. Based on previous research, social media data has been proven to store information related to customer and market conditions. Having the ability to process and analyze social media data can be a competitive advantage for companies. However, not all companies can carry out social media analysis. Lack of the ability to process data and high infrastructure costs are among the causes. One of them happens to SMEs in Indonesia, where they have minimal resources. We propose a solution in the form of a social media analytics platform to retrieve, process, and analyze social media data and make it valuable information based on these problems. We use multi-label text classification techniques in text mining to identify content on social media posts and group them into business decision-making metrics. This platform is expected to help small and medium enterprises make more appropriate business strategy decisions.

Keywords: Social Media Analytics, Text Mining, Business Decision, Small and Medium Enterprises

Introduction

The existence of social media encourages the growth of user-generated content. User-generated content is created and shared by people on the internet in text, images, or videos. User-generated content in social media posts summarizes customer wants and needs for a product or service [1]. With appropriate analytical methods, the information stored in these social media posts content can be extracted and used for various business purposes, especially to formulate business strategies. However, not all companies can take advantage of this social media data. Many small and medium enterprises are still not utilizing social media data for business purposes due to the lack of capabilities and resources. One of them is for small and medium enterprises / SMEs in Indonesia. SMEs in Indonesia still face problems accessing market information, which ultimately results in a lack of ability to make the right strategy in marketing their products [2].

Based on these problems, we propose a solution in the form of a social media analytics platform that can extract, analyze, and summarize social media data into business dimensions such as marketing mix, service quality dimensions, and product quality dimensions. This social media analytics platform is expected to help SMEs in Indonesia get up-to-date market insights that lead to better decision making and increased competitiveness of Indonesian SMEs.

Literature Review

a. Social Media Analytics

Social Media Analytics is a technique for extracting valuable information from social media data [3]. Social media analytics enables insightful decision making based on user-generated content on social media. Social media analytics has been widely applied in the business sector to provide market-related knowledge or answer business problems [3].

*SBM-ITB, muhammad-apriandito@sbm-itb.ac.id

b. Text Mining

Text mining is a method for extracting insights from data in text form [4]. Text mining is closely related to Natural Language Processing, which studies how computers can understand human language. Text mining has been widely used to extract information in text data on social media and use it for business purposes [5].

This study uses the Multi-Label Text Classification technique, which aims to group text into one or more predetermined dimensions. Text data will be grouped into business dimensions, namely the marketing mix, service quality dimensions, and product quality dimensions.

c. Marketing Mix

The marketing mix is defined as a combination of various variables in marketing decisions used by a business organization to market its goods and services [6]. The purpose of the marketing mix is to decide on a strategy to meet customer needs. The marketing mix is known by the four Ps elements; namely, Product, Price, Place, Promotion explained in table 2.1. The company controls the 4Ps marketing mix to influence buyer responses. This mix is essential to consider by every manager and business owner to achieve a competitive advantage for the organization.

Table 2.1 Marketing Mix Elements

Elements	Explanation
Product	Physical products or services that will be paid for by consumers. This includes both tangible goods and intangible products (services).
Price	The nominal amount that must be paid by consumers to get the offer.
Place	Commonly known as distribution channels, they can be physical or virtual stores.
Promotion	Communicate and persuade the target market to buy the product offered.

d. Service Quality

Quality in the user-based approach is defined as "fitness for use" and in a product-based approach is defined as “a precise and measurable variable” [7]. The service quality was developed as an advanced model for measuring service quality. There are five measurement dimensions in this model which are shown in the table 2.2. This model is formed through empirical studies of service quality in various service fields.

Table 2.1 Marketing Mix Elements

Dimensions	**Explanation**
Tangibles	Equipment, physical facilities, and personal appearance
Reliability	Perform the promised services accurately and reliably
Responsiveness	Willingness to help customers and provide prompt service
Guarantee	The ability of employees to inspire trust and confidence, knowledge of manners
Empathy	Caring, customer understanding, easy access to individual attention that a company gives to its customers

Methodology

There are three steps taken to create a social media analysis platform, as shown in Figure

3.1. The first stage is to define the flow of data processing and analytics, the second stage is to package the data processing and analysis flow that has been made into a web platform, and the last stage is to deploy the web platform so that it can be accessed publicly via the internet.

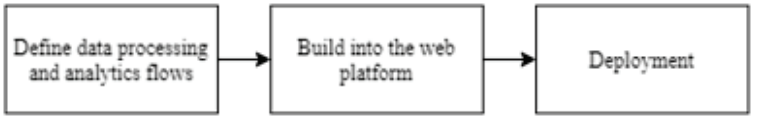


Figure 3.1 Stage of creating a social media analytics platform

In the first stage, create a data processing flow. We define all the processes that will be carried out, starting from collecting to visualizing. There are four processes: collection, preprocessing, analyzing, and visualization.

We use Twitter as a data source because Twitter still provides an API that can be accessed for data collection, so the data obtained is legal. After the data is obtained, the tweet content data is in the form of text, and then preprocessing is carried out to improve the quality of the data by adjusting the structure for data analysis. The preprocessing process is carried out by removing symbols, transforming the case to lowercase, tokenizing data into a token, reducing a word to its word stem, and stop word removal. Furthermore, we also do word weighting using Term-Frequency - Inverse Document Frequency (TF-IDF) weighting.

After the preprocessing process, the text data were analyzed. In this section, we first create an analysis model using the Multi-Label Text Classification Technique to classify tweets

into business dimensions. The classification algorithm used is the Naive Bayes algorithm. Based on previous research, Naive Bayes has proven capable of classifying text [5]. The business

dimensions that are used are the Marketing Mix and Service Quality.

After all the analysis processes have been made, the next step is to package it into a web platform. We use the Shiny package in the R programming language. There are two components needed to create a web platform using Shiny, namely UI and Server. The UI defines how the platform looks, and the Server defines how the shiny platform works. In the UI section, we created a UI where users can interact, including input keywords and data collection parameters. In addition, we also define the location where the output will be displayed. In the Server section, we load data processing and the analytics model, then make a connection between the UI and the Server section.

After the social media analytics platform has been developed, the next step is deployment. The deployment stage aims to make this social media analytics platform publicly accessible via the internet. We use R studio's shinyapp.io service to deploy the social media analysis platform that has been created.

Result and Discussion

After the application is deployed and publicly accessible, we try to analyze the public conversations on Twitter regarding "Haircut" in Indonesian. In the keyword input section, we entered the word "Haircut" in Bahasa Indonesia and set the data collection parameters, as shown in Figure 4.1.

Gambar 4.1 Input keyword and data collection parameters

After pressing the "analysis" button, the social media analysis platform will automatically retrieve the tweet data based on the inputted keywords, analyze it based on the predetermined dimensions, and visualize tweets' frequency in each dimension as in figure 4.2, which shows the results of the analysis of the dimensions of the marketing mix, and service quality.

Figure 4.2 The analysis result of the marketing mix and service quality dimensions related to "Haircut"

As can be seen in Figure 4.2, the most discussed dimensions are the "Place" and "Product" dimensions. While in the dimension of service quality, the most discussed dimensions are "Tangibility" and "Reliability". This result suggests that when people talk about "haircuts," they talk about the barbershop condition and the barber.

Conclusion

We have successfully created a business-related social media analysis platform for SMEs in Indonesia to identify market desires and help make business decisions. Currently, this platform is in testing with selected SMEs. In future research, we hope to add more business dimensions, such as PESTLE and product quality dimensions, to enrich the insights generated by this social media analytics platform.

References

[1] *Alamsyah, A., Saputra, M., & Masrury, R. (2019). Object Detection Using Convolutional Neural Network To Identify Popular Fashion Product. Journal of Physics Conference Series.*

[2] *Khan, G. F. (2015). Seven layers of social media analytics: Mining business insights from social media ; text, actions, networks, hyperlinks, apps, search engine, and location data.*

Erscheinungsort nicht ermittelbar: Createspace.

[3] *Taufiq, Rahmat & Jatmika, Dwi & Kunci, Kata & Ukm, & Usaha, & Jatmika, Rahmat. (2017). Masalah yang dihadapi Usaha Kecil Menengah di Indonesia.*

[4] *Liu, B. (2013). Web data mining: Exploring hyperlinks, contents, and usage data. Berlin: Springer.*

[5] *Saputra, M., Alamsyah, A., & Fatihan, F. (2020). Hotel preference rank based on online customer review. Test Engineering and Management.*

[6] *Kotler, P., Armstrong, G., & Opresnik, M. O. (2021). Principles of marketing. Harlow, England: Pearson.*

[7] *Yarimoglu, Emel. (2014). A Review on Dimensions of Service Quality Models. Journal of Marketing Management. 2. 79-93.*

A review paper on Identification of the CAPTCHA with the advancement of Machine Learning Techniques

Surendra Kumar Pathak*
Dr. Naveen Kr Singh**
R.K. Maurya***

ABSTRACT

Captcha is very useful as it is differentiate between human and machine. We enter the specified text in the given box, so that it can differentiate between human and BOT. It is assumed that bot will not recognize the captcha text because it is provided in a difficult fashion. With the development of the Machine Learning techniques, new and powerful algorithms have been developed, this leads to the break of the Captcha text. It also arise the chance that the bot automatically recognize the text and enter it into the provided space.In this paper we will discuss various issues due to the advancement of the ML on Captcha.

Keywords: *CAPTCHA, Machine Learning, Bot, Convolution Neural Network (CNN).*

Introduction

CAPTCHA is not a new word for the web users. In our day to day life we encounter with this terminology. CAPTCHA (Completely Automated Public Turing Test to tell Computers and Human Apart) – is a computer program which is used to differentiate between computer and human. It is a program which is used as a bot to keep away the other bots. In simple word, it keeps the other machine or program which can automatically access one's machine, since machines are not capable of identifying distract images as the human does.

According to Ved Prakash Singh et al, there are different types of CAPTCHA available:

- Text based CAPTCHA
- Mathematical CAPTCHA
- Image based CAPTCHA
- Audio based CAPTCHA and
- Re- CAPTCHA

Figure 1 Different Captcha images

**Assistant Professor, Dept. of Computer Applications, JIMS Engineering Management Technical Campus, Greater Noida, Surendra.gn@jagannath.org*
***Professor, Information Technology, IPEM, Ghaziabad, drnaveenkrsingh@gmail.com*
****Associate Professor, Deptt .of Computer Application, ABESEC Ghaziabad.*

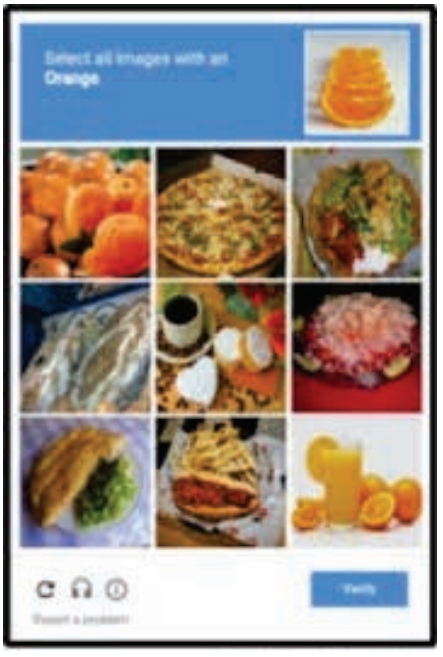


Figure 2 Image based Captcha

There are some techniques available to solve CAPTCHA. Some of the techniques are:

- Online Captcha solving services
- Optical Character Recognition(OCR)
- Machine Learning

In this paper we will focus only on Machine Learning Techniques.

Literature Review

Captcha was introduced in the year 2000 and as we have entered in the year 2020, we should see that Captcha will be capable to handle the different bots which are now so powerful due to the advancement of Machine Learning.

According to Bostik and Klecka with the help of supervised learning techniques we can easily break the optical character recognition(OCR) of captcha codes. Now we have started using automated Captcha to solve different problems, some are illicit while others are using it legitimately to know its power.

Machine Learning is the branch of the Artificial Intelligence (AI), with the advancement of the processing power of the computers and availability of the huge data and the big data tools we are now able to access the captcha with the help of the ML programs.We know, a lots of data is generating everydata. No matter whether it is structured, unstructured or semi structured data.However quantity of generating the unstructured data is very large as it is generating through Facebook, twitter, whatsaap and other social media platforms.Availability of huge data and enormous computing power provides us better use of Machine Learning techniques.

Now ML is matured enough and it provides the solution of different textual and voice recognition problems with the help of Natural Language Processing (NLP).

Methodology

With the help of the Convolution Neural Network (CNN), Computer Vision and different Python libraries such as tensor flow, we can provide training

to the machime using deep CNN models, to find the letters and digits in the image provided by the CAPTCHA.

OpenCV is used to locate contours in an image which detects the continuous regions. One can use thresholding to preprocess the images. All the captured images can be converted into black and white. With the help of the findContour() method of OpenCV we can break the Captcha images into letters and digits. The output of the processed image are now only letters and digits which is fed to the CNN for training. After training, the trained CNN model is ready to solve the real Captchas.

Conclusion: ML based CAPTCHA provides better solution than OCR however online services can provide better solution since machine learning is based on prediction and how accurate the prediction depends on the training and testing of the particular data set.

In a nutshell, we want to say thatwith the advancement of Machine Learning and availability of different Python libraries, Capcha can be identified. So we should think one step ahead to deal with this issue.

Bibliography

- *Naomi S. Altman. An Introduction to Kernel and NearestNeighbor Nonparametric Regression. Am. Stat., 46(3):175–185, aug 1992. ISSN 0003-1305. doi: 10.1080/00031305.1992.10475879.*
- *Elie Bursztein, Matthieu Martin, and John C. Mitchell.Text-based CAPTCHA strengths and weaknesses. Proc.18th ACM Conf. Comput. Commun. Secur., 2011:125–138, 2011. ISSN 15437221. doi: 10.1145/2046707.2046724.*
- *Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N.Vapnik. A training algorithm for optimal margin classifiers. In Proc. fifth Annu. Work. Comput. Learn. theory -COLT '92, pages 144–152, New York, New York, USA,*
- *Sagarmay Deb and Yanchun Zhang. An Overview ofContent-based Image Retrieval Techniques. 18th Int.Conf. Adv. Inf. Netw. Appl. 2004 AINA 2004, 1:59–64,2004. doi: 10.1109/AINA.2004.1283888.*
- *Davidek. Bubble Captcha - A Start of the New Directionof Text Captcha Scheme Development. In Mendel 2017,23rd Int. Conf. Soft Comput., volume 23 of 23, pages57–64. Brno University of Technology, 2017.*
- *Xiao Ling-Zi and ZHANG Yi-Chun "A Case Study of Text-BasedCAPTCHA Attacks," in International Conference on Cyber-EnabledDistributed Computing and Knowledge Discover, 2012.*
- *Carnegie Mellon University. The Official CAPTCHA Site,2010. URL http://www.captcha.net/.*
- *Corinna Cortes and Vladimir N. Vapnik. Support-vectornetworks. Mach. Learn., 20(3):273–297, sep 1995. doi:10.1007/BF00994018.*
- *Christopher M. Bishop. Pattern recognition and machinelearning. Springer, 2006. ISBN 0387310738.1992. ACM Press. ISBN 089791497X. doi: 10.1145/130385.130401.Ondrej Bostik, Karel Horak, Jan Klecka, and Daniel.*
- *Wei-Bin Lee, Che-Wei Fan ,Kevin Ho, Chyi-Ren Dow , and "ACAPTCHA with Tips Related to Alphabets Upper or Lower Case,"in Seventh International Conference on Broadband, Communication, Wireless Computing and Applications, 2012.*
- *Baljit Singh Saini and Anju Bala "A Review of Bot Protection usingCAPTCHA for Web Security," IOSR Journal of Computer Engineering, 2013, pp. 36-42, 2013.*

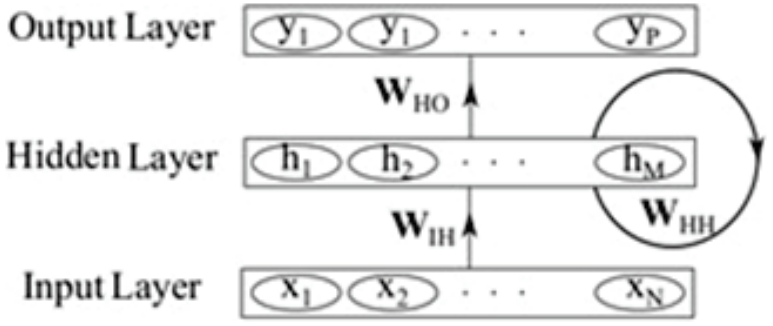


Fig. 1 Recurrent Neural Network(RNN)

Bahari, T. Femina, and M. Sudheep Elayidom[2] considers a CRM-data mining framework that establishes close customer relationships and manages relationship between organizations and customers in today's advanced world of businesses. Data mining has gained popularity in various CRM applica- tions in recent years and classification model is an important data mining technique useful in the field. The model is used to predict the behaviour of customers to enhance the decision-making processes for retaining valued cus- tomers. An efficient CRM-data mining framework is proposed in this paper and two classification models, Neural Networks and Naive Bayes are studied to show that the accuracy of Neural Network is comparatively better. To il- lustrate the performance of classification models the CRM applications such as customer segmentation, prospecting and acquisition, affinity and cross sell, profitability, retention and attrition, risk analyses, etc in banking domain are considered. Two classification models, the Multilayer Perception Neural Net- work (MLPNN) which have their roots in the artificial intelligence and Naive Bayes (NB) classifier, a simple probabilistic classifier based on applying Bayes' theorem are used for the study.

Surendro and Kridanto [5] uses predictive analyses in their paper. In the modern era of computing, organizations are focusing on the better utilization of technology and surviving to gear-up with global business demand. Such competition is acting as a driving force for its business to cope-up the data which generated every second of minute. This data needs to figure out and information which is required for the business growth model. The Predictive Analytics (PA) uses various algorithms to find out different patterns in large data that might suggest the efficient behavior for business solution. Today, different technologies can together transform the information technology but in turn, are imposing new complexities to the data computation. Due to such advances in technologies, and it requires rapid and dynamic data analysis for structured and unstructured data.

Zheng, Bichen, et al. [6] in the paper, customers' restaurant preferences are predicted based on social media location checks-ins. Historical preferences of the customer and the influence of the customers' social network are used in combination with the customers' mobility characteristics as inputs to the model. Artificial Neural Network and support vector machine are used to pre- dict the customers behavior on restaurant preferences. The framework of the propensity model is shown in Fig 2. The raw data, after pre-processing is con- verted into four key features: Predicting Time, Historical preferences, Friends' Recent Impact, and Transience Status.

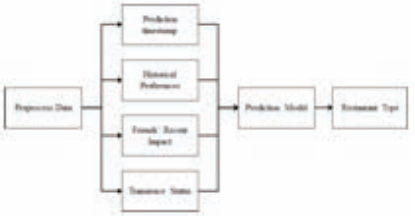


Fig. 2 Framework of prediction model

Customer Behaviour Prediction using Propensity Model

Remya K Sasi • Hima John • Binny Jerard •
Sujay Sudheer • Ashik Shaju

Dr. Remya K Sasi*
Hima John**

ABSTRACT

The Consumer Behavior Prediction model is a system which uses a combination of the techniques, machine learning and data mining to retrieve customers' retail behavior from the provided historical data, ie; this model is made to learn the history of purchases of an individual customer and from that a pattern is formulated(like the monthly households purchased, the amount of each item, the frequency, etc.) from a retail shop, say a supermarket. From this, the model is capable of predicting the possible future purchase of that customer/ consumer. This paper contains the literature survey done by our group and the analysis of the papers discussed so as to choose from the different techniques the suitable technique combination to implement our propensity model.

Introduction

Customized services is a crucial factor in retail markets. Customized services has become the key issue in developing the customer relationship. Service providers can maintain a long-term and a pleasant relationship with the cus- tomer if they can study the customer behaviour. This paper discusses state of art methods published in the area of customer behaviour prediction.

There are existing systems which are most commonly used to assess the customer behaviour online. This helps to customize user or customer expe- rience more effectively. Taking this concept as a guideline we are building a system that can provide efficient analysis on customer behaviour pattern in retail market, say a supermarket. The system analyses customers' purchase behaviour patterns and learn from it. In the end the system is supposed to pre- dict customers future purchases by making use of the historical data provided. The papers discussed below are sources of supporting analyses, experiments, hypotheses on customers retail behaviour and possible techniques to analyse the purchase patterns.

Literature Review

Salehinejad, Hojjat, and Shahryar[1]proposed a Deep learning technique that increases the efficiency of marketing strategies. The big data when analysed with deep learning techniques such as Recurrent Neural Network(RNN) pre- cisely predicts the customer behaviour. Here, the variables considered are client loyalty number (CLN), recency, frequency, and monetary (RFM). With these the customer behaviour patterns are predicted, ie the no of times the product is purchased, its time interval etc. With the use of Recurrent Neural Net- works, the customer data is refined in every steps of the network by feeding the output again as input to the next network layer. Fig. 1 shows a Simple Recurrent Network(SRN) which comprises of a input layer, hidden layer, and output layer. In the hidden layer, the weights are updated frequently so that an authentic result is predicted.

*Associate Professor, CSED, Christ College of Engineering, Irinjalakuda E-mail: remyaksasi@cce.edu.in
**Associate Professor, CSED, Christ College of Engineering, Irinjalakuda E-mail: remyaksasi@cce.edu.in

The first approach used for prediction is SVM and the second approach used is ANN. This methodology is based on the data pre-processing results ,traditional multi-layer perceptron neural networks and are trained by back- propagation to learn the relationship between the inputs and outputs.

Chen, Mu-Chen, Ai-Lun Chiu, and Hsu-Hwa Chang [7] uses data mining techniques for customers' behavior prediction. Data mining techniques search through a database for informations including knowledge rules, constraints and regularities.

Fig. 3 Data Mining

Association rule is initially used for customers' beviour patterns. Changes in customers' behavior are identified by comparing two sets of association rules generated from two datasets of different periods. Useful variables are hidden in a large quantity of raw data, and are obtained through data integration and transformation. Customers are segmented into various target markets in terms of customer value obtained by scoring RFM.In this study, the behavioral variables, RFM, coupled with growth matrix of customer value, are applied to estimate the value that individual customers contribute to the business. Finally, an online query system provides marketing managers a tool for rapid information search, and valuable information based on prompt feedback. The system enables marketing managers to rapidly establish marketing strategies.

Peker, Serhat, Altan Kocyigit, and P. Erhan Eren [8] paper has introduced the individual-level and the segment-based predictive modeling approaches. The purpose of this paper is to propose a hybrid approach which predicts customers' individual purchase behaviors. The proposed hybrid approach is established based on individual-level and segmentbased approaches and utilizes the historical transactional data and predictive algorithms to generate predictions. The effectiveness of the proposed approach is experimentally evaluated in the domain of supermarket shopping by using real-world data and using five popular machine learning classification algorithms including logistic regression, decision trees, support vector machines, neural networks and random forests. This approach substantially outperforms the individual-level and the segment- based approaches in terms of prediction coverage while maintaining roughly comparable prediction accuracy to the individual-level method. Moreover, the experimental results demonstrate that logistic regression performs better than the other classifiers in predicting customer purchase behavior. It concludes that the proposed approach would be beneficial for enterprises in terms of de- signing customized services and one-to-one marketing strategies. It is the first attempt to adopt a hybrid approach combining individual-level and segment- based approaches to predict customers' individual purchase behaviors.

Kulkarni et al. [11] tries to use the particular dimensions to quantify the complexity of customer in-store movements, and proposes a purchase model factors in the effects of complex customer movements on purchase behavior. We used the box-counting method to calculate the particular dimension of shopping paths and investigated its relationships with cart size and sales, which are viewed as important for marketing. We found that the customer group with high dimensions had mean values for the number of times visited on the online website, stay time in online store, and sales amount statistically higher than those of the customer group with lower dimensions. We analyzed a binomial logit model to identify positive effects that the dimension has on their purchases.

Damian, Andrei, et al [12] present a deep hierarchical recurrent encoder- decoder architecture that makes possible to account for sequences of previous queries of arbitrary lengths. As a result, our suggestions are sensitive to the order of queries in the context while avoiding data sparsity. Additionally, this model can suggest for rare, or long-tail, queries. The produced suggestions are synthetic and are sampled one word at a time, using computationally cheap decoding techniques. This is in contrast to current synthetic suggestion models relying upon machine learning pipelines and hand-engineered feature sets. Re- sults show that this model outperforms existing context-aware approaches in a next query prediction setting. In addition to query suggestion, the architecture is general enough to be used in a variety of other applications.

Motivation

When coming to purchases, a customer always goes with the trust. i.e; its a common trend seen in every consumer to stick to one trusted supplier. Everything and anything is available online these days. This always arise the question of trust. Even the seller is certified, a consumer with no prior experience with a dealer will be reluctant to make an order. Even with a prior experience its not easy for a consumer when it comes to online shopping. More or less the experience do have a factor of luck. This is the trigger to propose such a system which always help the trusted suppliers of the customers' to make their service even more effective and customized. when stating suppliers we refer to hypermarkets and supermarkets. As these are still the most trusted centers when coming to purchase. These suppliers by using this system can predict the monthly needs of its customer and verify with customer and deliver the order to their doorsteps. Door deliveries are no new thing to people these days. Therefore the acceptance of this project in the society can be set to high hopes. Now considering the pandemic situation which the country is going through, the importance of such a system is hiked to a certain level. Keep- ing social distance and do visit these outlets is still risky upto certain level. Thus, automating the consumer's purchases will not only reduce the chances of getting infected but also adds to convenience. After all its all about safety and trust. Experimental results show that the proposed model do just apt in solving the problem under consideration.

We are familiar with enterprises giving customer oriented services. This trend is widely followed in e-commerce platforms where they studies the customer/consumer behaviour to provide better, quality services. For example, these portals are able to give suggestions or recommendations according to our previous purchase history or search histories. The thought of applying this to retail stores give birth to this project. We expect to customise the services provided by these outlets. retail outlets like super markets and hyper markets are considered here. Those customers who are registered and are regular will be getting the privilege of customised service. This project is basically a pre- diction model that takes in the customer purchase history, then analyse it and predict the monthly necessary items they regularly purchases and notifies the customer beforehand at the month end for the purchase confirmation or add on. The system of door delivery is incorporated into this system so as to let the customer enjoy the smooth flow of busy life.

Background studies

Recurrent Neural network

A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Derived from feedfor- ward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. The term "recurrent neural network" is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic behavior. Recurrent neural networks were based on David Rumelhart's work in 1986. Both finite impulse and infinite impulse recurrent networks can have additional stored states, and the storage can be under direct control by the neural network.

RFM Model

RFM (Recency, Frequency, Monetary) analysis is a marketing technique used to determine quantitatively which customers are the best ones by

examining, how recently a customer has purchased (recency), how often they purchase (frequency), and how much the customer spends (monetary). RFM analysis is based on the marketing axiom that "80 percent of your business comes from 20 percent of your customers."

In customer behavior analysis, the available data in gross form are trans- actions relating to purchases made by the customers of a store. However, to make sense of this data and to be able to train learning models above they must be treated well to extract variables that identify behavior. In this con- text, We study the characteristics R, F and M (the recency, the frequency and the quantity of money spent by an individual on a given date), we answer three questions about the client: When did he buy for the last time? How many times has he bought? And how much does he have paid?

Objective

- The purpose of the above literature survey is to find out the possible com- binations of techniques or technologies that can be used in designing the model. A propensity model should be able to predict customers' purchase behaviour like whether the customer is likely to buy something.
- The customer purchase behaviour propensity model is a system with pre- diction techniques. This model when input with the purchase history data or information of customer on his/her retail market behaviour will output a pattern which is that customer's future purchase choices. Analysing differ- ent pre-existing models or system we hope to find out the best suited tech- nique combination to implement the Customer purchase behaviour propen- sity model.

Methodology

Consumer Behaviour Prediction Model is a propensity model which can predict the customers' purchase behaviour making use of the RFM analysis. This model is a deep learning system which is to be implemented in the retail market hubs like supermarkets. The data or the purchase history of customers, are used to analyse their pattern of purchase and to predict their future purchases. This is to provide and ensure the satisfactory in customer services provided which is a key in good and profitable marketing.

System Architecture

A propensity model for customers' behaviour prediction that uses the cus- tomers' purchase history to predict the future purchases of the customers. we are taking retail shops like supermarket under consideration where the cus- tomers hold an account. The history of each customer can be obtained by using this account. A pattern in purchases is formed assessing the frequency of purchases. Then the list of possible items that the customer may likely to buy is predicted using learning algorithms. The data source is from the online and offline customer databases. The data is pre-processed so as to clear the anomalies and to structure the data. Thereafter fed to the predicting algo- rithm to get the customer purchase prediction. The desired items are send to the customers as required.

Approach

The study of the parameters R, FandM and their evolution over time makes it possible to characterize the relationship with a client.The objective of this project is therefore is to implement a classification algorithm of customers based on their transaction history.The algorithm will be trained with data captured in time sequences T0, T1, ..., Tn to predict the class at a date Tn. Recurring model entries are sequential and take into account the evolution of the parameters R, F and M .

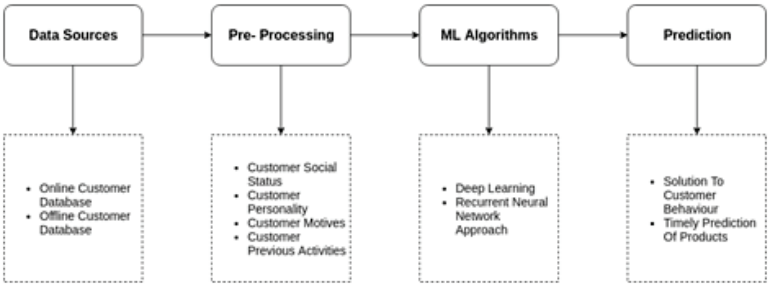


Fig. 4 Steps in Customer Behaviour Prediction

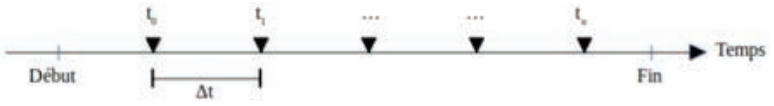


Fig. 5 Figure showing segmentation of the study interval

Dataset

The dataset used here is the purchase history of customers. This dataset is used as the training data. For this the amount of data we are considering is over the span of 5 years. This is the initial state . When this training is over the model is inputed with test data of the same format that of the training data. The performance is analysed and the training data is improved with adding more and more data to it. A sample dataset is given below:

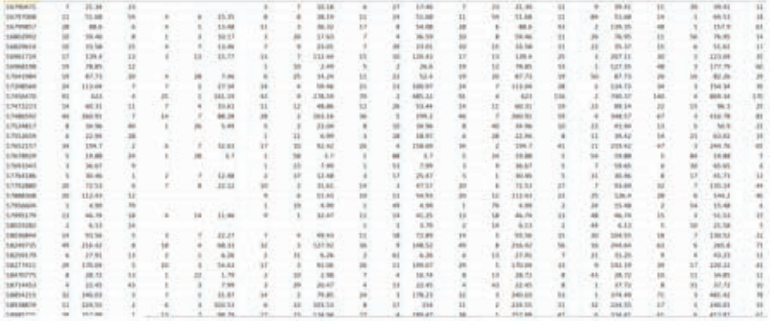


Fig. 6 Snap of sample

For a customer i:

- Rij the difference between the last transaction made and tj.
- Fij the number of transactions up to tj
- Mij the amount spent until tj

These data are then stored in the form of a tensor of dimension (n; p; q) which will be used to supply the algorithm.

Implementation

In order to study the behavior of customers and their changes over time we will define over the study period time intervals of equal size T (depending on the problem studied subdivisions can be daily, weekly, monthly). We then iterate on these subdivisions by considering each time the transactions carried out before a date tj and we calculate the Rij, Fij and Mij values for each customer i.

n: number of Dataset customers
p: number of parameters to enter (here 3: R;)
q: number of intervals in the study period: E = tn1t0T
The first n1 blocks are used to train the model. The values of R, F and M to Tn are used to define the class membership of each client. Indeed, according to these values clients are classified into groups i, j, k where i; j; k2[1 :: 4]. As an indication, an inventory

value high means that the customer has not been seen for almost a long time, so it is classified in category 4 for R. On the other hand, high values of F, M are appreciated, they reflect customer loyalty and significant expenses, they are in categories 4 for F and M.

Conclusion

We can achieve significantly better, accurate results when machine learning and data mining are combined. From the comparison study on the above papers reveal that the percentage difference in the combined results of these techniques is efficient and better in the case of machine learning and data mining. So we have decided to implement our Customers' purchase behaviour propensity model using machine learning and data mining techniques.

References

1. *Salehinejad, Hojjat, and Shahryar Rahnamayan. "Customer shopping pattern prediction: A recurrent neural network approach." 2016 IEEE Symposium Se- ries on Computational Intelligence (SSCI). IEEE, 2016.*
2. *Bahari, T. Femina, and M. Sudheep Elayidom. "An efficient CRM-data min- ing framework for the prediction of customer behaviour." Procedia computer science 46 (2015): 725-731. Author, Book title, page numbers. Publisher, place (year)*
3. *Wang, Xishun, Minjie Zhang, and Fenghui Ren. "Learning Customer Behaviors for Effective Load Forecasting." IEEE Transactions on Knowledge and Data Engineering 31.5 (2018): 938-951.*
4. *Valecha, Harsh, et al. "Prediction of Consumer Behaviour using Random Forest Algorithm." 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2018.*
5. *Surendro, Kridanto. "Predictive Analytics for Predicting Customer Behavior." 2019 International Conference of Artificial Intelligence and Information Tech- nology (ICAIIT). IEEE, 2019.*
6. *Zheng, Bichen, et al. "Customers' behavior prediction using artificial neural network." IIE Annual Conference. Proceedings. Institute of Industrial and Sys- tems Engineers (IISE), 2013.*
7. *Chen, Mu-Chen, Ai-Lun Chiu, and Hsu-Hwa Chang. "Mining changes in cus- tomer behavior in retail marketing." Expert Systems with Applications 28.4 (2005): 773-781.*
8. *Peker, Serhat, Altan Kocyigit, and P. Erhan Eren. "An empirical comparison of customer behavior modeling approaches for shopping list prediction." 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2018.*
9. *Liu, Di, et al. "Analysis and Accurate Prediction of User's Response Behavior in Incentive-Based Demand Response." IEEE Access 7 (2018): 3170-3180.*
10. *Chu, Yunghui, Hui-Kuo Yang, and Wen-Chih Peng. "Predicting Online User Purchase Behavior Based on Browsing History." 2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2019.*
11. *Kulkarni, Hrishikesh, Pramod Patil, and Radhika Menon. "Multi-Agent Sys- tem for Customer Behavior Tracking Using Shoppers' Path or Traversal." 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT). IEEE, 2019.*
12. *Damian, Andrei, et al. "Advanced Customer Activity Prediction Based on Deep Hierarchic Encoder-Decoders." 2019 22nd International Conference on Control Systems and Computer Science (CSCS). IEEE, 2019.*

Estimate of Modeling Units for Hindi Speech Recognition using Artificial Intelligence

Arjun Kumar*
Achintya Kr. Pandey**
Sudarshan Singh***

ABSTRACT

A speech recognizer is a digital system, which is very complex and developed to understand human speech. Its applications are used in different domains like computer narration and interaction with handheld devices such as cell phones, through to speaker independent tasks such as indexing videos for better search and automatic subtitling. The paper includes the development of a true Hindi speech recognition system which recognizes the isolated words as well as connected sequences of word for predefined fixed sized vocabulary and gives the output in Hindi text and a comparison of feature extraction techniques such as LPCC, MFCC and PLP are done.

Keywords: *LPCC, MFCC, PLP*

Introduction

In last fifty years, different speech recognition techniques have been given proposed and implemented. These techniques span many sciences, including pattern recognition, artificial intelligence, statistics, probability theory, information theory, computer algorithms, linguistics, psychology, and even biology. During this time, Automatic speech recognition has been matured markedly. This is happened because of the increment in computing power, and because of better modeling techniques. The statistical framework for ASR and introduction of the HMM in the 1970s[1], has been proven the most important successful approach to date, and worked as basis for current state-of-the-art speech recognizers. At initial phase, their main concentration was on isolated word recognition for small vocabularies, like the task of digit recognition. After getting positive results on previous tasks, their focus shifted towards continuous speech recognition of small and medium vocabulary. The performance on small and medium vocabulary has steadily improved and consequently. From last few years, research has begun to consider large vocabulary continuous speech recognition (LVCSR). But despite all these advances there is no such ASR system available in the market for Indian language with high accuracy. In the past, a significant portion of their search has been done on continuous, large vocabulary, speaker independent, speech recognition systems for language like English, Japanese and other European languages. These languages have got mature ASR engines by now. But Indian language engines are not mature enough and hence most current research is trying to improve the accuracy of the ASR engines in these languages [2,3]. There is still a long way to go before we will have an ASR system at par compare with engines for English.

Implementation of Hindi Speech Recognition System

This section describes the implementation of Hindi speech recognition system and comparative study of various feature extraction techniques. The systems have been developed on HTK-Toolkit V 3.4 in Linux environment (Ubuntu 10.04.3 LTS). Speech recognition system has been developed for 101 word vocabulary size for the Hindi language. Each word is

*Assistant Professor, Department of Information Technology, GNIT, Greater Noida, 9ansingh@gmail.com
**Assistant Professor, Deptt. of Computer Science & Engineering, GNIT, Greater Noida, achintyacs07@gmail.com
***Assistant Professor, Deptt. of Computer Science & Engineering, GNIT, Greater Noida, achintyacs07@gmail.com

uttered for a number of times to capture all the acoustic variability. The system has been developed in two parts namely front-end and back-end. Back-end covers acoustic modeling, recognition and language modeling while front-end part works on preprocessing and feature extraction while back-end covers acoustic modeling, language modeling and recognition. The chapter also presents a comparison of feature extraction techniques. The comparative analysis shows that MFCC perform better in same training and testing conditions while PLP perform better in mismatch conditions while both the feature extraction techniques outperform LPCC.

System Architecture

The developed speech system is divided into two parts: front-end module and back- end module. Data preparation is most important task and initially, it is done. All the words of the vocabulary are uttered a number of times. Since speech recognition system cannot directly processed speech sound because an acoustic signal is an analog signal. Acoustic signal should be represented in a much efficient and compact form which can be achieved by using acoustic analysis. Back-End module generates the system model which is to be used during testing.

Data Preparation

To implement a speech recognition system, a basic requirement is speech and text corpus. In this implementation works elf developed speech and text corpus issued. A unidirectional Sony microphone of 120VA issued for the preparation of speech corpus. Data is collected using 4 people (3males, 1female). Recording is done using system command brec. The properties of data are: sample is taken at sampling rate of 16000 Hz, bitrate16-bitandthe file format is PCM.wav. HTK also supports .sig file format, but it has compatibility issues. This file format is only supported by HTK, while .wav file format is supported by many other recognition tools. Data is prepared for limited vocabulary size of 101 words. Each word is uttered for ten times, so that speech corpus can capture most of the acoustic variability's. Text corpus is prepared manually using wave surfer. It takes lots of human hours but manually prepared speech and text corpus produce better results, if it is prepared with proper precautions.

Vocabulary: The system is developed on a limited size vocabulary. The system is developed for 101words. System performs well for any combination of these words for making sentences.

Experimental analysis of results:

The features were compared in general field conditions and in clear environment with known and unknown speakers. where known speakers are referred as data samples are recorded in corpus while unknown speakers are referred systems do not have their samples).Table presents the recognition results in terms of total spoken words and recognize word. The recognition results shows that the MFCC is better when testing and training conditions are same but in mismatch condition PLP out performs MFCC, while the performance of MFCC and PLP are better that LPCC in general field conditions and clean environment.

Table 1: Comparison of Results of Feature Extraction techniques

Feature Extraction Technique	No of Words	Speaker 1 Known	Speaker 2 Known	Speaker 3 Unknown	Speaker 4 Unknown
		50	55	60	65
LPCC	Clear	46	53	51	56
	Field Condition	45	52	48	51
MFCC	Clear	49	55	52	56
	Field Condition	46	52	49	51
PLP	Clear	48	54	51	55
	Field Condition	48	53	50	53

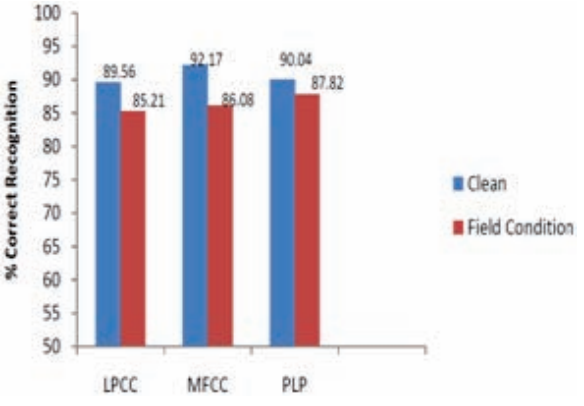


Figure1: Percentage of Correct Recognition

The recognition result uses all three feature extraction techniques; the speech recognition system uses same testing data to produce the result for all the three feature extraction techniques. The experimental result shows that MFCC feature extraction technique produces more accurate results.

The test data was prepared separately by a set of speakers and it was used to test the system. And speakers were asked to utter few words from vocabulary. From the training data some data was collected for testing purpose. Four speakers were selected to collect the test data. Out of these four speakers, data of two speakers were used for training the system. Therefore for testing the system three types of sounds were used: sound used to trained the system, sound of speaker who's different sound files were used to trained the system, and sounds of a unknown speaker whose sounds does not used for training purpose. Recognition results in Figure show that LPCC, MFCC, PLP produces89.56%, 92.17%, 90.04%respectively correct recognition. While in general field conditions the percentage of correct word recognition is respectively 85.21, 86.08 and 87.82.

Conclusions

The goal of Hindi ASR system is that the system should recognize the acoustic signal and give the output transcription in Hindi. The limitation of previous developed systems is that the transcription of acoustic signal comes in English. The contribution of work is to develop a true Hindi speech recognition system in which transcription of recognize words comes in Hindi. The system was developed for vocabulary size of 101words. To develop the system, Hidden Markov model tool kit (HTK) was used that uses Hidden Markov models (HMM) for recognition. MFCCs are used to extract acoustic features. It was calculated for the experimental results, the developed system produces an accuracy of 92.17% in clean environment and 86.08%in general field conditions. The experimental results of feature comparison show that MFCC is better when testing and training conditions are same but in mismatch condition PLP out performs MFCC, while the performance of MFCC and PLP are better that LPCC in general field conditions and clean environment. It was also found that the system is producing better results with more vocabulary- size as compared to similar works done by others.

References

1. *L.E.Baumand, J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology", Bull .Amer. Math. Soc., vol. 73, pp. 360-363, 1967.*
2. *M. Kumar, A. Verma and N. Rajput, "A large vocabulary speech recognition system for hindi," Journal of IBM Research, vol. 48, pp. 703-715, 2004.*
3. *K. Samudravijaya, "Computer recognition of spoken Hindi", Tata Institute of Fundamental research, 2000.*
4. *N. Chomsky and M. Halle, "The sound pattern of english", MIT Press, Cambridge, MA, 1991.*
5. *A. Rosenberg, L. Rabiner, J. Wilpon and D. Kahn, " Demi syllable-based isolated word recognition system", Acoustics, Speech and Signal Processing, IEEE Transactions, vol. 31, no. 3, pp. 713 -726, 1983.*
6. *J. Allen, M. S. Hunnicutt and D. Klatt, "From text to speech: The MI Talk system" Cambridge Studies in Speech and Science Communication, Cambridge University Press, Cambridge, 1987.*
7. *M. A. Randolph, "A data-driven method for discovering and predicting allophonic variation", IEEE Conference on Acoustics, Speech, and Signal Processing, pp. 1177 - 1180, 1990.*
8. *V. Zue, "The use of phonetic rules in automatic speech recognition", Speech Communication, vol. 2, pp. 181–186, 1983.*
9. *K. W. Church, "A stochastic parts program and no unphrase parser for unrestricted text", IEEE Conference on Acoustics, Speech, and Signal Processing, pp. 695-698, 1989.*
10. *F. Jelinek, L. Bahl and R. Mercer, "Design of a linguistic statistical decoder for the recognition of continuous speech", IEEE Transactions on Information Theory,vol.21, no. 3, pp. 250 -256, 1975.*
11. *J. L. Gauvain, L. F. Lamel, G. Adda and M.D.Adda," Speaker independent continuous speech dictation", In: Proc. EURO SPEECH, Berlin, Germany, pp. 125–128, 1993.*
12. *J. R. Glass and T. J. Hazen, "Telephone-based conversational speech recognition in the Jupiter domain" In: ICSLP, Sydney, Australia, pp.1327–1330, 1998.*

13. *M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin and D. Bell, "Linguistic constraints in hidden Markov model based speech recognition", In: Proc. ICASSP, Glasgow, Scotland, pp. 699–702, 1989.*
14. *P. Fetter, A. Kaltenmeier, T. Kuhn and P. Regel, "Improved modeling of OOV words in spontaneous speech", In Proceedings IEEE international conference on acoustics, speech, and signal processing, vol. 1, pp. 534–537, 1996.*
15. *S. Young, "A review of large vocabulary continuous speech recognition", IEEE Signal Processing Magazine, vol. 13, 45–57, 1996.*
16. *C. H. Lee, J. L. Gauvain, R. Pieraccini and L. R. Rabiner, "Large vocabulary speech recognition using sub word units", Speech Communication, vol. 13, 263–279, 1993.*
17. *P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev and S. J. Young, "The development of the HTK large vocabulary speech recognition system", InProc.ARPA spoken language systems technology workshop, pp. 104–109, 1995.*
18. *J. K. Baker, "Stochastic modeling for automatic speech recognition", in Speech Recognition, edited by D. R. Reddy, Academic Press, 1975.*
19. *F. Jelinek, "Continuous speech recognition by statistical methods", Proceedings of the IEEE, vol. 64 no. 4, pp. 532- 557, 1976.*
20. *A. Poritz, "Hidden Markov Models: A Guided Tour", Proc. Of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1. pp. 1-4, 1988.*
21. *A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from in complete data via the EM algorithm", Journals of the Royal Statistical Society, vol.39, no. 1, pp.1-21, 1987.*
22. *L. Baum, "An inequality and associated maximization technique occurring in statistical estimation for probabilistic functions of a Markov process", Inequalities, vol.3,pp.1-8, 1972.*
23. *L. Breiman, J. Friedman, R. Olshen and C. Stone, Classification and Regression Trees, Wadsworth & Brooks, Pacific Grove, CA, 1984.*
24. *L. Bahletal., "Estimating hidden Markov model parameters so as to maximize speech recognition accuracy",IEEE Trans. Speech and Audio Processing,vol.5,no.2,pp.179- 190, 1993.*
25. *R. Lippman, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, vol. 4, no. 2, pp. 4-22, 1987.*
26. *F. Beaufays, H. Bourlard, H. Franco and N. Morgan, "Speech Recognition Technology," In Handbook of Brain Theory and Neural Networks, 2nd edition, M. Arbibed., MIT Press, 2002.*
27. *A. L. Buchsbaum, and R. Giancarlo, "Algorithmic aspects in speech recognition: An introduction", Journal of Experimental Algorithmic, vol. 2, no. 1, 1997.*

Crypto currency: A Futuristic Digital Currency for the Digital World

Dr. Gavendra Singh*
Mr. Afendi Abdi**
Mr. RajaSekhar Boddu***
Mr. Adugna Alemayehu****

ABSTRACT

In October 2008, after the fall down of Lehman Brothers, A Pseudo name, and Anonymous Developer, SATOSHI NAKAMOTO published a paper on the crypto currency known as BITCOIN based on the concept of Proof of work that stop double spending. On January 3, 2009, Satoshi Nakamato mined the very first Bit coin, that's Called "genesis block."Satoshi Nakamoto combined several prior inventions such as b-money and Hash Cash to create a completely de-centralized electronic cash system that does not rely on a central authority for currency issuance or settlement and validation of transactions.However, it's not clear that Satoshi Nakamoto is an individual or a group of peoples who wrote the security proof paper about technology of Bit coin.

Keyword: *Crypto currency, Bit coin Block chain, cryptography*

Introduction

Crypto currency is an open source, peer to peer, decentralized (i.e. no centralized server), public distributed, ledger-based technology.The mathematics involved are impressive, and the use of specialized hardware to construct this vast chain of cryptographic data renders it practically impossible to replicate. So crypto currency is online technology payment system without the involvement of consultancy fee of third party.[1][2]

Literature Review

If we will read in the literature of the finance, we can easily observe that those payments methods are limited in quantity but they are increasing in their value like all the precious metals (i.e. Gold, silver). Gold and silver were the main components of payments method, initially people didn't like this idea of using preciousmetal for exchange of goods but later it was accepted and became famous for food and other goods exchange. Later it became difficult to mine the gold and silver currency.[3] Afterthis, paper currency was introduced in the countries and it also became popular.Butwe can observe that paper currency value is decreases if Govt's print more paper money. So, in case of Bit coin it's a limited in terms of space and block.

Some of the Earlier Digital or Virtual Currency

E-Gold

One of the first virtual currencies was E-gold, founded in 1996. E-gold was uniquein that its virtual currency was backed by real, honest-to-goodness gold bullion [3]

Beenz and Flooz

In 1998, an interesting new website called Beenz.com was launched. The ideabehind Beenz.com is that you could earn virtual currency (called Beenz) for performinga variety of online activities, such as

**Assistant Professor, Department of Software Engineering, College of Computing & Informatics, Haramaya University, P.O. Box 138 Dire Dawa, Ethiopia, yashgaven11@gmail.com*
***HOD, Department of Software Engineering, College of Computing & Informatics, Haramaya University, P.O. Box 138 Dire Dawa, Ethiopia, afe2003@gmail.com*
****Lecturer, Department of Software Engineering, College of Computing & Informatics, Haramaya University, P.O. Box 138 Dire Dawa, Ethiopia, rajsekhar.cse@gmail.com*
*****Lecturer, Department of Software Engineering, College of Computing & Informatics, Haramaya University, P.O. Box 138 Dire Dawa, Ethiopia, adugna10@gmail.com*

visiting certain websites or shoppingonline. The Beenz you earned could then be spent on various online goods andservices.The site tried to position itself as "the web's currency" that would challenge theworld's traditional currencies [2][4].

COIN

The Chinese Internet service provider Tencent has a very successful instant messaging service called QQ. Back in 2002, QQ developed its own internal virtual currency, called Q Coins, that customers could use to purchase various virtual goods and services, such as extra storage space, virtual pets, and online game avatars [4].

Linden Dollars

The concept of virtual currency makes a lot of sense within online virtual worlds.

Case in point, the virtual world of Second Life and its very popular virtual currency, LindenDollars [5].

FaceBook Credits

In 2009 Facebook began testing the concept ofFacebook Credits, which could be used to pay for in-game goods and services onthe Facebook site. Facebook Credits went live in January 2011, and users couldpurchase 10 Facebook Credits for one U.S. dollar [6].

Relationship between Cryptography and crypto currency

"Crypto" means concealed or secret or anonymous. Cryptography technology ensures pseudo- or full anonymity. The cryptographic technology guarantees the security of the transactions by verifying the users, independence from the Central authority, and protection from spending the concept of Cryptography isused inthe Crypto currencies.

The first one is Symmetric Encryption Cryptography. It uses the same secret key to encrypt the raw message at source, transmit the encrypted message to the recipient, and then decrypt the message at the destination. A simple example is representing alphabets with numbers – say, 'A' is '01', 'B' is '02', and so on. A message like "AFENDI" will be encrypted as "010605140409" and this value will be transmitted over the network to the recipient(s). Once received, the recipient will decrypt it using the same reverse methodology – '08' is 'H', '05' is 'E', and so on, to get the original message value "AFENDI." Even if unauthorized usersare trying to read the encrypted message "010605140409," it will be of no use to them unless they know the encryption key and methodology. This is one of the basic and simplest examples of symmetric encryption, but lots of complex encryption cryptography exist for enhanced security.

The second method is Asymmetric Encryption Cryptography, which uses two different keys – public and private – to encrypt and decrypt data. The public key can be distributed openly, but private key is known only to the owner. In this technique, a sender can encrypt a message using the receiver's public key, but it can be decrypted only by the receiver's private key. This method helps achieve the two important functions of authentication and encryption for crypto currency transactions.

The third cryptography method is hashing, which is used to efficiently verify the integrity of data of transactions on the network. It maintains the structure of block chain data, encodes people's account addresses, is an integral part of the process of encrypting transactions that occur between accounts, and makes block mining possible. Transaction "blocks" are signed with a digital signature using a private key, Additionally, Digital Signatures complement these various cryptography processes, by allowing genuine participants to prove their identities to the network. Multiple variations of the above methods with desired levels of customization can be implemented across various crypto currencynetworks [7].

Bit coin and other crypto currency based on SHA-256Algorithm

SHA-256 stands for "Secure Hash Algorithm" that belongs to SHA-2 family. It generates 256 bit (32 byte) signature for a text string. Its block processing time is generally around seven minutes and requires hash rates at giga hashes per second. This mining algorithm was used by bit coin, the most popular crypto currency. Some of the other crypto currencies that are mined using this algorithm are Bit coin Cash, Dev Coin, Peer coin, Steem Dollars, Terracoin, Tiger coinetc. Other Popular crypto currencies Based On

The Concept Of block chain Bitcoin (SHA256), Ethereum (ETHASH), Lite Coin (SCRYPT), Monero (CRYPTONIGHT), Dash(X11), Zcash (EQUIHASH), Stellar, Tron Cardano. [8]

RIPPLE: Unlike Bit coin or Ethereum, Ripple doesn't have a block chain. A crypto currency without a Block chain may sound pretty strange - if it doesn't have a Block chain, how does it verify transactions and makes sure everything is ok? For that purpose Ripple has its own patented technology: the Ripple protocol consensus algorithm (RPCA).[9]

Main Components of Crypto currency

1. **Blockchain** is a particular type or subset of so-called distributed ledger technology ("DLT"). 8 DLT is a way of recording and sharing data across multiple data stores (also known as ledgers), each have the exact same data records and are collectively maintained and controlled by a distributed network of computer servers, which are called nodes. Block chain is a mechanism that employs an encryption method known as cryptography and uses (a set of) specific mathematical algorithms to create and verify a continuously growing data structure – to which data can only be added and from which existing data cannot be removed – that takes the form of a chain of "transaction blocks", which functions as a distributed ledger.

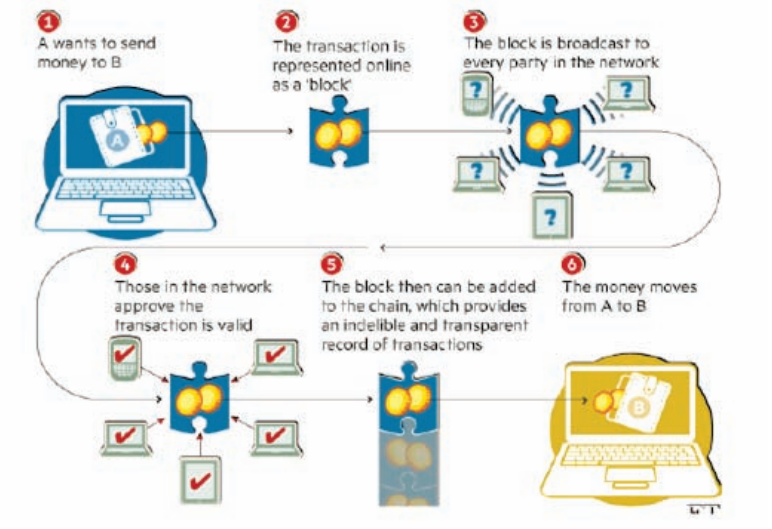


Figure 2: Working of Blockchain

Source: "Technology: Banks seeks the key to blockchain", by J. Wild, M. Arnold and P. Stafford, 1 November 2015, Financial Times, https://www.ft.com/content/

2. **Cryptocurrencyuser:** a first, and very importantplayer is the "crypto currency user". A crypto currency user is a natural person or legal entity who obtains coins to use them (i) to purchase real or virtual goods or services (from a set of specific merchants), (ii) to make P2P payments, or (iii) to hold them for investment purposes (i.e. in a speculative manner).

3. **Mineris** a player who participates in validating transactions on the block chain by solving a "cryptographic puzzle". As explained above, the process of mining relates to crypto currencies that are based on a POW consensus mechanism.

4. POW (Proof of work)

In a PoW system, network participants have to solve so-called "cryptographic puzzles" to be allowed to add new "blocks" to the blockchain. This puzzle-solving process is commonly referred to as "mining".

5. **Cryptocurrency** exchanges Another group of key players are the so-called "cryptocurrency exchanges". Cryptocurrency exchanges are persons or entities who offer exchange services to cryptocurrency users, usually against payment of a certain fee (i.e. a commission). They allow cryptocurrency users to sell their coins for fiat currency or buy new coins with fiat currency. They usually function both as a bourse and as a form of exchange office. Examples of well-known cryptocurrency exchanges are: Bitfinex91, Coinbase GDAX

6. **Wallet** providers Another group of key players are the so-called "wallet providers". Wallet providers are those entities that provide cryptocurrency users digital wallets or e-wallets which are used for holding, storing and transferring

7. **Coin** inventors There are also players who are referred to as "coin inventors". Coin inventors are individuals or organizations who have developed the technical foundations of a cryptocurrency and set the initial rules for its use.

Requirements of Effective and Efficient Hardware for Mining Bitcoin

Good Bit coin mining hardware should have a high hash rate. But, efficiency is just as important. An efficient Bit coin miner means that you pay less in electricity costs per hash. For Effective Bit coinMining we need a minimum CPU equivalent to an Intel Pentium 4 2.00GHz. However, the developers recommend a CPU greater or equal to an Intel Core i5-3330 for Bit coin Mining. You should have at least an NVIDIA GeForce GT 430 graphics card. Furthermore, an AMD Radeon HD 6950 is recommended in order to run Bit coin Mining. Bit coinmining system requirements state that you will need at least 4 GB of RAM. If possible, make sure you have 8 GB of RAM in order to run Bit coinmining task to its full potential.Bit coin Mining System will run on PC system with Windows 7 64bit and above version.

ASIC (Application Specific Integrated Circuits) Bit coin Miner

Since it's now impossible to profitably mine Bit coin with your computer, you'll need specialized hardware called ASICs. Here's what an ASIC miner looks like up close: Originally, Bit coin's creator intended for Bit coin to be mined on CPUs (your laptop or desktop computer). However, Bit coin miners discovered they could get more hashing power from graphic cards. Graphic cards were then surpassed by Asics. Think of a Bit coin ASIC as specialized Bit coin mining computers, Bit coin mining machines, or "bit coingenerators". Now a day's all serious Bit coin mining is performed on dedicated Bit coin mining hardware ASICs, usually in thermally-regulated data-centers with low-cost electricity.

Figure 2:DRAGONMINT 16T MINER.

Conclusion

In this paper we may conclude that these digital crypto currencies are open source and security proof software projects. The main advantage of these crypto currencies is that they don't have central authority and consequently no one is in a position to make fraudulent. Since the near future, we may see that more organization's might use the digital crypto currency to increase in their profits, so crypto currency may control the market, but at the same time govt's across the globe have to ensure that these crypto currencies should not be used by the black traders ,drug landers or for the use of black money.

References

[1] *Nakamato S., Bit coin: A peer to peer electronic cash system:2008. //https: bitcoin.org*

[2] *https://cointelegraph.com/bitcoin-for-beginners/what-are-cryptocurrencies*

[3] *https://www.moneycrashers.com/cryptocurrency-history-bitcoin-alternatives*

[4] *Coinmarketcap, Crypto-Currency Market Capitalizations; 2016. Accessed: 24/3/2016. https://coinmarketcap.com/.*

[5] *https://en.bitcoin.it/wiki/Linden_Dollars*

[6] *https://www.worldcryptoindex.com/how-cryptography-is-used-cryptocurrency*

[7] *Faulkner, J., Getting started with Cryptography in .NET, München BookRix, 2016, 121p.*

[8] *GRINBERG, R., "Bitcoin: An Innovative Alternative Digital Currency", Hastings Science & Technology Law Journal, 2011, Vol. 4, 50p. (electronically available via https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1817857).*

[9] *https://ripple.com/xrp/*

[10] *https://en.bitcoin.it/wiki/Proof_of_Stake.*

[11] *Antonopoulos AM. Mastering Bitcoin: unlocking digital cryptocurrencies. ª O'Reilly Media, Inc.º; 2014.*

[12] *GLAZER, P., "An Overview of Privacy Coins", February 2018, https://hackernoon.com/an-overview-of-privacy-tokens-19f6af8077b7.*

[13] *GOLDBERG, S., "Mythbusting: Blockchain and Cryptocurrencies Edition", May 2018, http://paymentsjournal.com/mythbusting-blockchain-and-cryptocurrencies-edition/.*

[14] *https://medium.com/supplyframe-hardware/the-impressive-hardware-used-in-cryptocurrency-mining-31771edb857c*

[15] *Ultimate guide for Bitcoin by Michael Miller,2015, USA*

[16] *https://en.bitcoin.it/wiki/Double-spending.*

Blockchain with Artificial Intelligence

Shrikant Patel*
Indu Jolly**
Girish Kumar Sharma***
Sanjay Kumar****

ABSTRACT

Artificial Intelligence and Blockchain are some popular concepts. To handle big data, Blockchain is a secure and transparent medium. AI, on the other hand, trying to take over conventional human mechanisms or clumsy algorithms and replace it with intelligent and smart coding that can learn from the data it collects. In this paper, we discuss how artificial intelligence and blockchain affect each other and the benefits of combining these two trends. Some use cases and companies implementing these concepts.

Keywords-*Blockchain, AI (Artificial Intelligence), Decentralized AI, Bitcoin, Swarn*

Introduction

Artificial Intelligence and Blockchain are topics of great interest. Both are infrastructure technologies historically comparable to desktop operating systems and communication among the internet. To understand the relation between these two first, we need to understand what Artificial Intelligence and blockchain are and how they are related to one another. Blockchain can initialize interactions between participants with no intermediary. AI, on the other hand, makes machines intelligent enough for decision-making similar to humans [1].

Artificial Intelligence: Intelligence specifies capability to make sense of information besides the normal as the name suggests data, knowledge that created artificially. Some of the AI techniques include robotics, machine learning. The goal of Artificial Intelligence is to make machines learn and apply the gathered data to become more intelligent. It is a way of making computer-controlled robot software thinking intelligently i.e. the way of giving computer ability to think and act smartly.

Blockchian: Blockchain is decentralized, and data management technology develops to solve the issue of currency transactions between persons or companies without involving third parties. Making a digital payment requires the involvement of the bank or the middleman to complete purchase or sale, they also charge a fee for the transaction thus making it centralized, and the third party controlled all the data instead of two-parties. As a survey, 8 of 10 people exploring block chain, whether investing in financial exploring blockchain industry or to make entirely new models for business. The results of the survey entitled Forbes "Blockchain as Blockbuster: Still Too Soon to Tell, But Get Ready" [2].

*Research Scholar, School of Computing Science and Engineering, Galgotias University, Greater Noida, U.P., India, patelshrikant@rediffmail.com
**Post Graduate Student, Bhai Parmanand Institute of Business Studies-Govt. Of Delhi, Delhi, India, indujolly171@gmail.com
***Professor, Bhai Parmanand Institute of Business Studies-Govt. Of Delhi, Delhi, India, gkps123@mail.com
****Professor, School of Computing Science and Engineering, Galgotias University, Greater Noida, U.P., India, drkatiyarsanjay@gmail.com

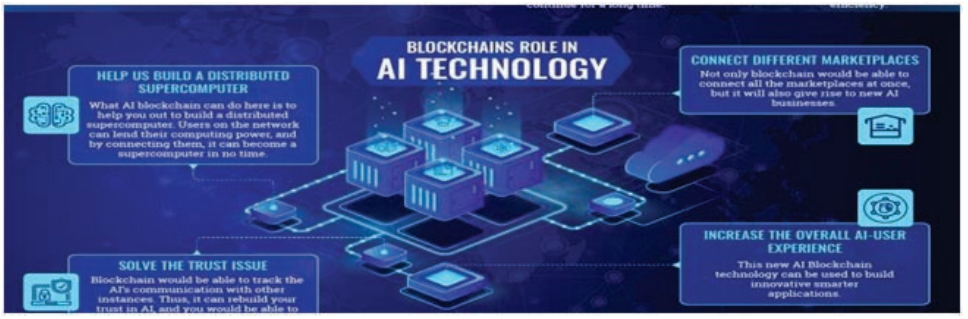


Figure 1. Why Blockchain is so popular

Artificial Intelligence and Blockchain Merger

As we know, blockchain relates to keeping accurate records, execution while A.I. relates to making decisions, understanding autonomous interaction. Both share a different feature, which leads to ensuring continuous communication in the future. Blockchain's importance can be evident by the number of crypto currencies which continuously growing and currently exceeding 1900[4]. Three key characteristics are:

1. Data Sharing: On a particular network, a decentralized database needs to share data between multiple parties. As we have more and more open data to study, therefore the prediction of Machines is considered precise and accurate, and more reliable algorithms are produced. While accessing data, it provides transparency and also encourages the sharing of data of different peripherals of AI including the data, algorithms, and performance i.e. computing power. Hence, blockchain successfully able to create a decentralized way for the marketplace [3].

2. **Security:** Security is necessary on the blockchain network, especially while handling transactions done via the pre-existing protocols. For AI, machines have autonomous nature, which requires high security to reduce the chances of an unfortunate occurrence.

3. **Trust:** Lack of trust is a danger to the enhancement of widely accepted technology and AI or blockchain both not excluded and to achieve machine to machine communication, a high level of trust is needed. While executing a particular transaction on the blockchain network, faith is required.

AI and Blockchain: An Integrated Approach

Want to implement Blockchain Technology AI can help you in this. The figure gives you an insight into merging Artificial Intelligence with blockchain.

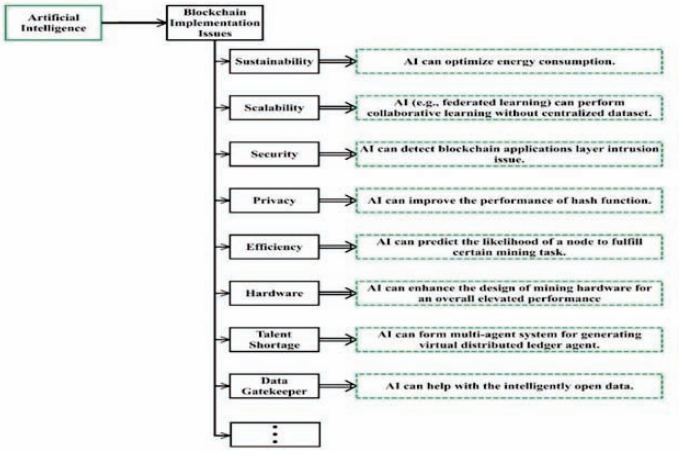


Figure 2. Merging block chain with AI

1. **Sustainability:** To analyze the microeconomic environment, basic tools like power system planning and operation, intelligent optimization algorithms are used to optimize the performance of a large-scale system by making use of AI. Both blockchain and microeconomics are distributed large-scale systems, inherently connecting each other having many similarities, like similar unified systems, decentralized computations, and much more.

2. **Scalability:** As the number of users increases the scalability of blockchain also increases. Practically speaking, some factors are affecting the effectiveness of a blockchain system such as the time required for initiating a transaction, the time consumed for validating a transaction, and cost of each established communication. Each block of the blockchain contains a particular quantity of transactional data; the situation can be handled by traditional centralized data mining techniques [5].

3. **Security:** As the blockchain is not viable to hack almost, its applications are not much secure, such as the Bitfinex, etc [4]. In the last two years, looking at machine learning progress, AI is suitable to make blockchain to guarantee security to applications. The security includes data encryption mechanisms etc. Regarding that computational intelligence also acts a crucial role in cryptographic systems. One of the prominent rewards of using the power of computational intelligence is developing more tough ciphers and to improve the attack-defense method of system, developing the blockchain system through computational intelligence can be beneficial.

4. **Privacy:** As the blockchain system includes personal data; data encryption becomes a major issue for ensuring the privacy of the user. Privacy relates to security issues in which AI plays a crucial role. For example, the Bitcoin blockchain system, which makes use of elliptic curves based on private and public key generation [5].

5. **Efficiency:** For the Blockchain, maintenance is not just taking the total throughput maximized.

6. **Performance:** To understand the performance of the desired transaction. Consider a sensor network, while using it to track some object the total throughput maximization may cause some issues among different nodes. AI performs active learning to boost overall system performance.

7. **Hardware:** To keep a blockchain system running, computer components play an important role. Currently, the architecture of the computer is based on Architecture given by Von Neumann according to it, a computer is consists of various machinery such as CPU, memory, storage, Input/Output devices, and buses (wires used for making connections). There are some other architectures of computer also like reduced instruction set computer, and parallel processing.

8. **Talent Shortage:** To decrease the limited contribution of the blockchain workforce, Multi-agent is a single approach that can be used. To automate the whole procedure of transaction data writing/reading from blocks virtual agent which are task-oriented can be used. To nurture the blockchain, AI technology can be used to a great extent.

9. **Data Gatekeeper:** As the popularity of the economy of data increases, intelligence and data are of main concern. Data resources based on blockchain become more accessible, users seek help with the data in their ease of understanding, consumption, and security. The highly powered technology of AI perfectly suits for this.

How Blockchain can Change AI

In the expansion of machine learning systems, Blockchain can help a lot. It can:

- Help in explaining AI better: There is an explanation problem with AI. It provides an obvious way to track back the machine decision

procedure, the reliability of the data as well as, of the models, [6].

- Used to increase the effectiveness of AI: To make data sharing secure means lots of data and need for enhanced models, improved actions, superior results.
- Lower the barriers of the market to entry: Blockchain technology can protect the information and organize the personal data. Second, it allows fresh marketplaces: a data market, a models market, and finally, an AI market. Hence data-sharing becomes reliable and to provide a better way that lowers the obstacle to entry for smaller new market-places with blockchain is data validation And data monetization.
- To increase trust in artificial intelligence: Tasks will manage by autonomous virtual machines. They also increase every peer to peer interaction and contract, establishing a mode to distribute data and robust mechanism securely.

Important Benefits of Blockchain With AI

The major benefit of AI is studying the patterns in a large amount of data, whereas blockchain has the main focus with precise records management and security. The sharing of data is the chief concern of AI and blockchain. Artificial intelligence connects with information and blockchain is a way of secure data transfer. A Self- operating device has autonomous nature hence need great inter-device communication; this is one of another problem that blockchains can crack. Blockchains make sure the validation of data across the internet. Models of Machine learning have "Trash In, Trash Out" nature– To develop the model, if any compromise over data is done, and then the model output won't be advantageous [7].

- AI and blockchain together: Blockchain holds highly secure data. Hence used for storing very sensitive, personal data which, when processed smartly then value information can be extracted. Information can only be added to preceding data and, after entering data cannot be modified, or lost [2]. For example, Amazon's or Netflix's recommendation engines to suggest what we might like to buy or watch. Obviously, the data feed into these systems is very personal and needs high data security. Information is stored in an encrypted form in database of blockchain, which means that only the private keys should be kept protected to secure all information on the chain. AI has lots to bring in terms of security. AI is linked with making algorithms that are able of dealing with information also in an encrypted state.

- Blockchain help us to trail, recognize and clarify AI-based decisions: Decisions prepared by AI can be tough for humans to realize because they are able of assessing a large number of variables autonomously of each other and knowledge, which one is important to the overall task. As an example, AI algorithms use in making decisions about whether monetary transactions are fraud and should be barren or investigated [8]. For example, Walmart feeds all its transactional data across each store into its AI systems, which make decisions on what goods should be stocked, and where [6]. AI can manage blockchains in a better way than humans. Traditionally, computers have been high-speed, but very stupid as they are unable to handle clear orders on how to perform a task. Working with blockchain data on networks needs large amounts of computer processing power.

Why Companies are Using Blockchain to Power AI

There can also be a data market that can be useful for developers looking for certain types of data for their projects. Although there is great interest in data to drive AI research, AI can boost the blockchain in many ways. AI provides an effective way to learn from history, while blockchain makes it possible to build trustworthy relationships by following a company's network. The combination of blockchain technology and artificial intelligence is still a largely unexplored area. Because blockchain is designed to obtain trustworthy data, this is the blood of artificial intelligence. Projects dedicated to this ground-breaking combination are rare, although the convergence of these two technologies has attracted considerable attention in both the public and private sectors, as well as in science and industry.

Data is one of the world's most valuable resources, and combining these two technologies has the

potential to use data in ways we never thought possible. Early last year, several companies focused on optimizing value chains announced plans to enter the blockchain scene - a small twist that has become a growing trend - by combining blockchain with the power of artificial intelligence (AI). Blockchain allows individuals to monetize the data they produce and allows us to review the intermediate steps that AI takes to conclude from that data. The move stems from the fact that blockchain, used by decentralized, independent parties, is, at its most basic level, a register of all information collected. Using the data stored on the blockchain, Hypersmart Contracts acts as a kind of AI-powered brain that can detect and solve complex efficiency and optimization problems, enabling the company to immediately release payments with crypto currencies. Computable Labs has developed a blockchain-based tool to build a data market for AI based applications. By building this AI market, it facilitates the creation of data and algorithms that the computable community can view, buy, and sell. How blockchain is used: How does it work and how can you use it in your own business?

Computable is developing a dedicated marketplace for health care, civil society, business, and industry to collect and share data with accredited members of the community, which will increase efficiency. Finalize aims to streamline this critical process and maximize Region of Interest in an industry that is estimated to reach $16.5 trillion in revenue by 2027. Blackbox Artificial Intelligence builds artificial intelligence tools for new technologies and how it uses blockchain tools. The company's engineers have developed a bespoke information architecture based on blockchain that focuses on smart contracts and intelligent contract management

Blockchain can create trust and a degree of transparency to relieve such concerns, and Microsoft, which helps thousands of companies manage their data, claims it can. It is an ecosystem of data exchange and monetization that provides a tokenized level of service to enable access to data from AI-enabled companies such as cloud computing, cloud storage, and data analysis. The idea of integrating these technologies has attracted an enormous number of entrepreneurs and venture capitalists, as can be seen in the blockchain use cases - AI - mentioned above. Synapse uses AI blockchain convergence to create a decentralized trainer-researcher-processor contract that can be addressed in real-time and programmatically. However, the process of combining these two technologies is not as simple as it appears. The benefits of these technologies can be exploited through integration with other cutting-edge technologies. This synergy can cover most problem areas such as machine learning, artificial intelligence, and deep learning.

The integration of blockchain technology with IoT and AI has far-reaching implications. The above example is a supply chain in which the IoT measures a lot of different metrics about the environment during a journey, uses the blockchain to store the data, and then uses AI on that data to make human-like decisions. Another purpose of a blockchain solution is to create transparency by executing smart contracts [9].

Companies are looking to merge AI into their products and services to be a part of the competition are Enigma, Datum, Nucypher, ocean protocol, Computable labs. Involvement of some companies MICROSOFT assures that blockchain can append trust and transparency hence they developed "Azure Blockchain Data Manager". It is software that collects on chain data and connects it to other applications therefore business data can be sent from nodes (inside smart contracts) to other databases [12].

ORACLE provides cloud service for blockchain which is easy to put into practice, cost efficient while maintaining security. It is a business blockchain stage intended for enterprises and easy to organize and incorporate. It is carried with software development kit and API's for making integration ease, straightforward to include smart contracts to communications with third party, and also able to connect with existing applications [13].

Future of these Trends

Blockchain and Artificial Intelligence are hilarious technologies trending nowadays. Some of interesting trends are:

1. **SingularityNET:** Singularity Net is the storm to the world on December 2nd, z. marketplace where companies can purchase and sell A.I. algorithms, machine learning tools, and data sets on a worldwide scale leads to the expansion of the industry. SingularityNET makes AI accessible for all globally. In this stage, agents make use of AGI tokens to pay for AI correlated services. SingularityNET reflects that A.I. and blockchain both are future. SingularityNET trying to develop an advanced general intelligence that used for almost any task. SingularityNET able anyone to create, share AI services on a global scale.

2. **Namahe-A.I. Supply chain:** The aim of the majority of industries today is to raise automation for effectiveness and cost reduction either on the hardware part or on the software. Namahe tries to bring AI to blockchain supply chain management. Managing the supply chain's major goals in the field of cryptocurrency. Namahe is the first socially responsible supply chain. From aircraft parts to vegetables, the blockchain is affecting the shipping and warehousing industry on a large scale. Namahe merge AI and blockchain to make a protected environment, where businesses make savings on expensive audits and get better precision in their supply chains by utilizing of the AI level of Namahe [15].

3. Blockchain as a service (Baas): Various cloud

TRENDS	OBJECTIVE	APPLICATIONS	BLOCKCHAIN BENEFITS
Expandable AI	Designing trustworthy algorithms to know why algorithm is reaching at a particular decision	- Healthcare - Military - Autonomous vehicles	- Trust - Tracing executions - Reliability
Digital Twins	Translating data and intelligence from complex physical systems to applications in digital world	- wind turbines - aircraft engines	- Trust - derivation - Reliability
Automated Machine Learning	Automating the process of machine learning from raw data to manageable one in order to reduce human work and faster performance of applications	- Big data analytics - bulk production of Intelligent devices	- durability - Immutability
Hybrid learning models	Combining different machine learning models to get better decisions	- real time - decision-agonistic - data-agonistic	- Trust - derivation - performance
Lean and augmented data learning	Enabling transfer learning among various AI applications to ensure accuracy of data	- low data availability applications	- Trust - derivation - Reliability

Figure 3. Trends in different technologies

vendors such as Microsoft, IBM, and Amazon are developing their platforms to blockchain for their customers. Baas is one of the prominent technologies of blockchain which is integrated with startups and enterprises. Baas is just a cloud based service that enables users to make use of blockchain while developing their digital products such as smart contracts or any other service that works well without setting up and requirements of the entire blockchain network. Companies like Amazon, IBM, and Microsoft are building their blockchain network that provides Baas services to their customers.

Real Life Platforms

Some of amazing applications of blockchain and AI are:

1. **GAINFY (NEW YORK):** Gainfy is a platform for healthcare that makes use of blockchain, AI, and IoT to enhance industry standards. The company can indulge these technologies in the healthcare industry in many ways such as a database for clinical trials, a cryptocurrency payment system, an identity authentication, etc. Currently, investors receive tokens to enlarge service of gainfy in the healthcare.

2. **BLACKBIRD.AI (SAN FRANCISCO):** Blackbird AI rates the integrity of news content by making use of blockchain and AI. In struggling with false news, the corporation deploys Artificial Intelligence to categorize the data based on trustworthiness index such as identifying misleading information, spoof, and false speech. It behaves as a record keeping system that conclusively saves confirmed information produced by AI.

3. **AI BLOCKCHAIN (HOBOKEN, N.J):** AI Blockchain is a virtual record keeping system that arranges virtual agents to manage the sequence and blockchain of business work as a cyber security tool and digital record book for real estate, media, healthcare industry, and AI is used to manage the procedure of manufacturing and maintaining the blockchains [10]. AI Blockchain also acts as a stage for smart contracts due to consideration and reliable token rewards.

4. **BOTCHAIN (BOSTON):** Botchain enables worldwide registrations, audits, management of agents for AI software. AI actived bot log all information on the blockchain and receive botcoin by successively using applications over a dispersed network, and also bot is verified on the Botchain ecosystem. It is functioning in companies using AI in healthcare industry to bring their data keeping technology to bots to improve security and performance.

5. **DOPAMINE (NEW YORK):** Dopamine is a decentralized platform to monetize the intellectual property of AI and data providers. There is a dopamine community where information providers distribute their databases, and solutions base on AI and blockchain, if the solutions are reliable enough to solve a problem then they are rewarded. To log rewards and enumerate the company's reputation, company's blockchain maintain a record book on each data provider. Langnet is a company using dopamine to enlarge its AI information of the language, words and accent Langnet is an AI company focused on language data.

6. **SYNAPSE AI (SAN FRANCISCO):** Synapse AI is a network that gives rewards to the persons belongs to them with SYN tokens (crypto currency) for distribution of private information. Users can trade in synapse marketplace with syn tokens for bitcoin, etherum, USD by sharing their data. By sharing data in synapse marketplace also promote data buyers in promotion, pharmaceutical, learning industry to contact a large set of data. Currently, 100 million tokens are distributed to premature investors, and companies that purchase into its decentralized platform.

7. **COMPUTABLE (SAN FRANCISCO):** Computable Labs are popular for emergent tools based on that develop a decentralized marketplace for applications of AI. For the computable community, they facilitate a way to a broader cluster of data and algos that can be seen, purchased, and sold. Also the company is developing particular marketplaces for civic, healthcare, and business to gather and distribute statistics with the verified pervons in a society to boost effectiveness.

8. **NUMERAI (SAN FRANCISCO):** Numerai is a platform which is decentralized enclose fund at which a number of data scientists from all around the world are constantly functioning on problems of Blockchain and AI. It also features contest every week that run on machine learning and etherum for the data scientists in which they give raw statistics on machine learning troubles, scientists produce data model based on the data and the winners are rewarded by numerai (NMR) crypto. Every time scientists propose the model to it, they gave a piece of their own NMR crypto. If the model is chosen their crypto gets doubled or tripled of his bet.

9. **NEUREAL (SALT LAKE CITY):** Neureal combines AI, Cloud technologies, and blockchain to guess all from the supply market to searches on google. To predict future, AI examines earlier period prediction, and its blockchain record book logs every output so computer network can locate trends in exact predication. It works in various sectors.

Currently, the company is functioning on an AI based technique to predict the accurate course of storms.

10. **OBEN (PASADENA, CALIF):** Oben combines blockchain and AI to develop a stage for intelligent Artificial Reality and Virtual Reality avatars to craft a personal and social understanding in the implicit era. Users can produce their avatars that can interact with one another on Oben. Oben is teaming up with South Korean talent agency to develop the first AI agency for celebrities for making virtual celebrities as personal assistance or play the consumer's favourite song.

11. **LIVEEDU (SANTA MONICA, CALIF):** LiveEdu is an education platform which provides online lectures to students on how to develop real life products. Lectures are provided for AI design to directions on how to build crypto-currencies, and they use blockchain-based smart contracts in making payments to the content creators. Over 270,000 videos are available on the Liveedu platform on how to build up all from video games to AR tools, data science databases and AI robots.

12. **HANNAH SYSTEMS (SAN FRANCISCO):** To automate vehicles, Hannah systems combine AI and blockchain. The company also has a platform for data exchange which is AI based, tools for real time mapping, an imminent dashboard, and blockchain to take up, understand, and securely accumulate data in autonomous vehicles.

SWARN ROBOTICS: A Technology that benefits from Blockchain and Swarn Robotics is an area of great interest as it combines benefits of blockchain as well as Artificial Intelligence. In this area, various robots work together to execute tasks, and each robot able to interact with its surroundings by making use of AI by subsequent present rules. As a result, each of them is connected therefore their combined behaviour and interaction ability enhances.

While blockchain helps in security concerns by using advanced encrypted techniques like cryptographic digital signatures and secure public-key cryptography. Accessibility of the information is controlled by the private key provided to each robot. Hence, AI-powered robotics emerged as a front end technology and blockchain provides an optimal security solution [11].

Figure 4. Swarn Robotics: perfect use of blockchain and AI

Covid-19 as we all know is the greatest threat nowadays; it was declared a pandemic on January 30, 2020. It is an infectious disease caused by coronavirus and lots of research is going on it in order to find a way to fight with the virus. The government takes a lot of initiative to fight with the virus including lockdown. People infected with the virus feel mild to moderate respiratory problems and can recover by taking care of themselves with no need for special treatment. When a coronavirus positive person cough or sneeze, the virus spread through droplets of saliva or discharge of nose. Protect yourself from virus by washing your hands frequently, sanitize hands with alcohol based hand

rub and avoid touching your face. Last but not least boost your immunity by eating and drinking healthy food items.

Pandemic Covid-19 affects almost every sector adversely. Blockchain and AI help to study coronavirus data analytics, prediction, vaccine discovery. Blockchain can be used to gather patient's data more efficiently, monitor movements of the patient's to ensure social distancing.

- **Contact Tracing:** Across Europe, a team of privacy experts develops a COVID-19 contact tracing blockchain based system which is used to ensure privacy and data integrity by making use of the Decentralized privacy-preserving proximity tracing (DP-PPT). German tech MYXNG developed a blockchain based system that enables mobile phone tracing while taking care of data security. Governments and healthcare can get information about coronavirus tracking.

- **Supply Chain Management:** Many factories have been shut down due to safety issues which are exceptional demand for certain goods particularly medical supplies. Blockchain is suitable for supply chain management as it brings together all stakeholders and provides a source of data integrity. It also provides data transparency and ensures data security. Blockchain solution supplier TYMLEZ gives its services to Dutch govt and deploys solution that matches supply and needs in medical products. Mobile and online payment platform Alipay has developed a blockchain based solution to help charitable organisations work together more powerfully. It can track donations of relief supplies.

- **Finding Drugs:** The entire world desperately finding solutions to decrease the extent of the coronavirus and to find an efficient way of handling coronavirus. Technology is acting as a boon to formulate the procedure sooner. AI is performing a significant part in telling the mechanism of vaccine by considering viral protein structures and serving medical researchers polish tons of appropriate research at an extraordinary speed.

AI tools, shared data sets, and research results are created at Allen Institute for AI, Google deep mind [14]. In January, Google DeepMind developed an advanced system name "AlphaFold" that predicts the 3Dimensional arrangement of a protein-based on its inherited sequence. By using a popular technique of biology a three dimensional atomic map of element of virus that infects human cells–the spike protein is developed at the University of Texas, Austin and the National Institutes of Health. AI also assists scientists to discover a vaccine for coronavirus.

To enhance public health as well as help out in viewing people with exhibiting symptoms of illness and counsel them whether they need a treatment artificial intelligence can help. According to Pinaki Laskar, Founder and CEO, FishEyeBox, "We are being attacked by an 8 kilobytes virus and now living with the terms of coronavirus. AI can catch coronavirus symptoms before it spreads" [15]. The Artificial neural network (AAN) is used to guess the probability of the occurrence of disease. AI- enabled interfaces along with cameras can be used for face recognition, action recognition, to measure body temperature and consult to doctor which help people who are corona positive and using big data analytics including tracking on people's movement data and permit users to locate if they have approach in contact with a coronavirus person.

To assist organizations to track and envision coronavirus eruption, blockchain is a bonus. Acoer has build Hashlog dashboard which makes people know the level of growth of coronavirus. WHO (world health organisation) and Central Disease Center (CDC) used the blockchain hashlog dashboard to made a data visualization model associated with clinical trial as well as information and trends on social media. Also to develop a mutual system of sustainable, resourceful, and apparent drug supply chain, Blockchain plays a vital role in distribution application. Hence, Blockchain is advantageous that can help in the smooth functioning of medical supply chains making sure that drugs accessible in the market should not be bogus. In the current situation, the aim is to trace and track the medical supplies –whether they are genuine or standard.

The major areas in which blockchain works are liability, authority, and recognition of drugs. The data that blockchain generates from various sources, AI predicts the risk of distraction and forged drugs developed for the virus.

Figure 5. Use of these emerging technologies

Conclusion

The development of the Internet has evolved, and considering that artificial intelligence (AI) can now collect all data over a distributed network and extract meaningful information from it, this is happening faster than ever before. If we create a more efficient, secure, faster, easier to integrate, and more powerful network that benefits all of us, we will raise more questions than answers about who can benefit.

We are changing the foundations and fabric of the Internet, and as a result, IoT's are intelligent systems at all levels that define, design, and develop the infrastructure needed to connect, store, process, capture, and process information. As a result, we are creating a secure IoT ecosystem that is capable of performing tasks such as storing information and acting on time that no nation seems prepared for.

The paper describes blockchain and AI technologies, the benefits of these two fantastic technologies, and how AI can help in the combination of blockchain. Understand the merging capability and why to merge them, then some companies working in these technologies like computable labs and some trending technologies like Namahe which are future of blockchain and AI. Some real life applications based on these technologies such as Oben, Hannah system, Liveeud, and also discussed about benefit of AI and blockchain in robotics. How COVID-19 affect these technologies and how these technologies help in fighting with covid-19 in different ways.

References

[1] *K.Salah, M.H. Rehman, N. Nizamuddin, and A. Al Fuqaha "Blockchain for AI: Review and Open Research Challenges" IEEE Access PP (99) December 2018*

[2] *Spyros Makridakis, Antonis Polemitis, George Giaglis, and Soula LoucaClerk Maxwell, "Blockchain: The Next Breakthrough in the Rapid Progress of AI"Artificial Intelligence – Emerging Trends and Applications, Marco Antonio Aceves-Fernandez, IntechOpen DOI:10.572/intechopen. 75668 in November 2018*

[3] *Smriti, Saru Dhir, and Madhurima Hooda, "Possibilities at the Intersection of AI and Blockchain Technology", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278 3075, Volume-9 Issue-1S, November 2019*

[4] *FranCasino, Thomas K.Dasaklis, and Constantinos-pyPatsakis, "A systematic literature review of blockchain-based applications: Current status, classification, and open issues". Available:https://www.sciencedirect.com/science/article/pii/S0736585318306324*

[5] *Tshilidzi Marwala, and Bo Xing, "Blockchain and Artificial Intelligence" in arxiv.org at Cornell University in 2018. Available:https://arxiv.org/ftp/arxiv/papers/1802/1802.04451.pdf*

[6] *Bernard Marr, "Artificial Intelligence And Blockchain: 3 Major Benefits Of Combining These Two Mega-Trends" in Forbes in March 2018. Available:https://www.forbes.com/sites/bernardmarr/2018/03/02/artificial-intelligence-and-blockchain-3-major-benefits-of-combining-these-two-mega-trends/#4b6515c64b44*

[7] "How Blockchain Empowers AI" published in March 2018. Available: https://www.cio.com/article/3263810/how-blockchain-empowers-artificial-intelligence.html
[8] Inside Mobile Blog, "How blockchain empowers AI" Available:https://www.compassintelligence.com/blog/howblockchain-empowers-ai
[9] Bitcoin Exchange Guide News Team, "How blockchain empowers AI", published on November 16, 2017
[10] Sam Daley, "Tastier coffee, hurricane prediction and fighting the opioid crisis: 31 ways Blockchain & AI make a powerful pair" in builtin in April, 2020. Available:https://builtin.com/artificial-intelligence/blockchain-ai-examples
[11] Juned Ghanchi, "How blockchain and AI helps robotics technology" published in Robotics Business Review on November 2019. Available:https://www.roboticsbusinessreview.com/ai/how-blockchain-and-ai-can-help-robotics-technologies/
[12] Ian Allison, "Microsoft is using blockchain to help firm trust AI" published on Dec 2019 Available:https://www.coindesk.com/microsoft-isusing-blockchain-to-help-firms-trust-ai
[13] Oracle, "Transformation Technologies: Today" Available:http://www.oracle.com/us/solutions/cloud/tt-technologies-white-paper- 4498079.pdf
[14] ETCIO, "Covid-19: 8 ways in which technology helps Pandemic Management" published on April 2020. Available:https://cio.economictimes.indiatimes.com/news/nextgen-technologies/covid-19-8-ways-in- which-technology-helps-pandemic-management/75139759
[15] ET Government, "Coronavirus: Distruputive Technologies Network help to migrate future trends" published on March 2020. Available:https://government.economictimes.indiatimes.com/news/technology/coronavirus-disruptive-technologies-network-help-to-mitigate-future- threats/74668464

Investment Analysis based on the Principles of Value Investing using Machine Learning

Tanay Mehendale*
Anand Mane**
Vishal Ramina***
Shreyas Kailasnathan****

ABSTRACT

Security undoubtedly plays the main role of cloud CRM deployment, since the agile firms utilized cloud services in the provider infrastructures to perform acute CRM operations. In this paper we emphasis on the cloud CRM themes, benefits, security threads the most concern. Some aspects of security discussed concern on deployment the cloud CRM like: Access customers' database and control, secure data transfer over the cloud, trust among the enterprise and cloud service provider, confidentiality, integrity, availability triad, and security hazard, future studies and practice are presented at the end.

Keywords: *Cloud computing; CRM; Security; Cloud Security*

Introduction:

It is not possible for working professionals to invest the time needed to research a stock and come up with an informed decision. In this paper, the authors propose an investment analysis software which gives a signal as to whether a stock should be bought or not based on whether it is overvalued or undervalued. Deciding whether a stock is undervalued or overvalued is the base of value investing. The meaning and philosophy behind Value Investing is self explanatory. It is the process of finding the true value of a business and not deciding the value of a stock based on solely the market prices. It is a process where we look at a stock like a business and not just a piece contract whose supply and demand variations are taken advantage of in the short term. A listed stock is a business and when an investor buys a stock, he/she owns a part of the business and like an owner, makes decisions accordingly. The investor assesses the financial statements of a company thoroughly before investing and once he/she invests, the investor stays locked in for at least a couple of years to take full advantage of the situation. The proposed system makes an estimate on how much the stock is worth currently which is done by determining the intrinsic value of the company using Discounted Cash Flow model and Monte Carlo Simulation. It will then predict the market trend for the next 6 months using stacked Long Short Term Memory neural network.
Index Terms—Application programming interface (API), in- trinsic value, current market price, Discounted Cash Flow (DCF) model, Monte Carlo simulation, LSTM neural network

Introduction

Having discussed the importance of having a solid prin- cipled and disciplined approach to investing it is also very important to discuss the need for one to invest. No other security can give an investor the rate of return on an investment which a stock can give and systematic investing using proper risk management can help an individual achieve their financial goals much earlier than this in the same income bracket who don't invest in the stock market

*Department of Electronics and Telecommunication, Sardar Patel Institute of Technology, Mumbai, India, tanay.mehendale@spit.ac.in
**Department of Electronics and Telecommunication, Sardar Patel Institute of Technology, Mumbai, India, anand mane@spit.ac.in
***Department of Electronics and Telecommunication, Sardar Patel Institute of Technology, Mumbai, India, vishal.ramina@spit.ac.in
****Department of Electronics and Telecommunication, Sardar Patel Institute of Technology, Mumbai, India, shreyas.k@spit.ac.in

value. Out of all the different techniques used to invest in stocks, Value Investing is one that is time tested and has built wealth for every investor who stayed true to its principles, did not speculate and managed portfolio risk properly. Value Investing is the timeless technique to make money where in the investor buys a stock and holds it for a considerable period of time (for eg. 2-3 years on an average). The stocks to buy are the ones which are trading below the estimated intrinsic value. The intrinsic value is a very inexact value and can vary based on the interpretation of the person analysing the stock. It is what the investor thinks is the actual value of a business per share as opposed to what the market is quoting. This disparity between the market value and the intrinsic value is what drives an investment decision. Further focusing on the accuracy the proposed paper uses modern learning algorithms like the Monte Carlo simulation and the LSTM Network for prediction along with the classical timeless principles to make long term profits for investors. The use of financial data of the last 5 years for a particular stock makes sure that the model doesn't make the mistake of looking at the markets from a short term point of view. Beyond estimating whether the stock is undervalued or overvalued there is also an analysis done to compare the historical market value with the intrinsic value.

Related Work

While the approach towards the problem is novel, related works with respect to the problem statement we are addressing holds a significant value as the proposed paper would not have been possible without referring the following research:

The proposed paper incorporates Monte Carlo simulation [3] which is better than neural network when the uncertainty in the data is high. This approach is then modified in the proposed paper. The referred paper does not discount market inefficiencies in their analysis, which the proposed paper does using historical trading data before coming up with the final estimate. In [4] the concept of relative valuation is used using the partitioning around medoids method. The similarity of the stock in question is compared with the other stocks in that sector using the average in that sector as the reference. [6] gives a comparison of performance of five supervised machine learning algorithms. The five algorithms being Naive Bayes, Support Vector Machines, Random Forest, KNN and Softmax. However this paper analysed the data only for a maximum term of 90 days and the accuracy is less 58%. According to its conclusion Random Forest gives the best accuracy on large datasets which is tested in the proposed paper since it considers real time data as well. The paper [7] takes into account the case wherein we look to make up for the over reliance on book values. Though it is the key aspect of value investment which the proposed paper uses, it also looks at other parameters not mentioned on a balance sheet.

The above mentioned research works talk about either using financial ratios to calculate a certain parameter or using technologies in financial domain for gaining insight into a particular topic. However all of them, cannot be understood by a layman with little or no knowledge about financial markets. This was the motivation to propose this paper since this research aims to help individuals to invest in the market and also help existing investors irrespective of the level of expertise that they have in order to assess the company they are looking to invest in. No study work involved the effect of the current market trend on the company's intrinsic value. In addition to the valuation of the company, the prediction of how the company will perform in the next few months is also given in this paper.

Theory

Where all the focus is on making quick money with algorithmic trading and short term arbitrage strategies, the proposed paper looks to develop an investment analysis tool based on the classical principles of value investing. Taking into account the financial parameters of the last 5 financial years of a particular stock the proposed model estimates the intrinsic value of the company and judges whether it is undervalued and investment worthy using Monte Carlo simulation and LSTM neural network. Hence the proposed paper mixes some timeless principles of value investing with modern machine learning techniques and creates a framework where an individual with minimal knowledge of the financial markets can assess the company one is looking at and get a clear unbiased valuation.

A. APIs used
- Alpha Vantage
- Yahoo Finance

B. Methods
- DCF model and Monte Carlo Simulation
- Long Short Term Memory Neural Network

C. Parameters Considered

1) Price to Earnings Ratio: The Price-to-Earnings ratio (P/E ratio) is a proportion that allows all investors to assess the market value of the shares relative to the company's reported earnings reported on the three financial statements. In simple terms, the Price-to-Earnings ratio shows what it is that the management is currently willing to pay for the particular stock on its past and estimated performance.

The P/E ratio is essential as it provides a metric to compare if the stock is properly valued or not. A high P/E ratio may mean that the stock price is volatile compared to earnings and most likely overpriced. And a low Price-to-Earnings ratio could mean that the present share value is low priced in comparison to its earnings.

As the ratio indicates what an investor will have to pay for every dollar of return, a stock with such a lower Price-to- Earnings ratio ratio compared to the competitors in its sector pays less for every share for similar financial success than a stock with a greater Price-to-Earnings ratio. Value investors will use the P/E ratio to help identify undervalued stocks.

The price is taken for every day and divided by the earnings per share which we get from the balance sheet. A new column called p/e is formed which gives us the change in the ratio over the last one year.

2) Price-to-Book Ratio: The Price-to-Book ratio or P/B ratio determines whether the stock is over or underpriced by evaluating the total financial worth which is the difference between assets and liabilities of the company to its reported market capitalisation. Basically, the Price-to-Book ratio di- vides the share price of the stock by the book value per share (BVPS). The Price-to-Book ratio is a clear indicator of what investors are willing and able to pay for every dollar of the net value of the company.

A P/B ratio of 0.95, 1 or 1.1 implies that the underlying stock trades at almost a book value. In other words, the P/B ratio is much more beneficial the more the number varies from 1. For a profit-seeking investor, a company that operates for a P/B ratio of 0.5 is a great opportunity as it indicates that the stock value is half the worth of the company's book value.

The closing price per day is divided by the book value per share to give us the ratio of P/B which will help us in understanding the variation in the price.

3) Debt-to-Equity Ratio: The debt-to-equity ratio (D/E) is a stock parameter that helps investors determine how a company makes use of its current and non current assets. The ratio gives us the proportion of equity to debt a company is using to pay for its current and non current assets.

It is industry specific and can be a boon or a curse based on where the company is in its business cycle and how it has been doing in the recent past.

4) Free Cash Flow: Free cash flow (FCF) is the money generated by a business through its activities, minus the costs. In other words, the cash left over after a business pays for its capital expenditure (CapEx) is free cash flow.

Free cash flow can be an early indicator to value investors that earnings may increase in the future, since increasing free cash flow typically precedes increased earnings. when the earnings are low and the free cash flow is high it is a great sign for the company as high earning is round the corner.

5) Operating profits: We look at the profit after tax which is an important parameter in the proposed valuation process. The profit and free cash flow show a certain kind of correlation which speaks volumes about the company.

6) PEG Ratio: A updated version of the Price by earnings that also takes into account the growth in income from the income statement is the price/earnings-to-growth (PEG) ratio. The P/E ratio doesn't always give an estimate to an individual whether the ratio is suitable for the expected growth rate of the company or not.

A stock with a price/earnings-to-growth of less than 1 is usually considered undervalued because its price is low rel- ative to the anticipated earnings growth of the company. A price/earnings-to-growth greater than 1 could be deemed overvalued as it could mean that the stock price is too high relative to the projected earnings growth of the company. For

value investors, the PEG ratio is an important metric because it offers a future perspective to the business. These are all the parameters the proposed paper looks at. Now we look at how to model the data using machine learning and make an estimate of the intrinsic value and provide statistical analysis.

7) Monte Carlo Simulation and LSTM Network: All the parameters we discussed earlier will be analysed and financial statements will be fed as inputs to the Monte Carlo simulation after which we calculate the intrinsic value of the stock and determine whether it is worth investing in or not. It looks for undervalued stocks and also plots graphs and showcase other statistical inferences using these techniques. Future predictions of the stock trend will be done using LSTM network.

Proposed System

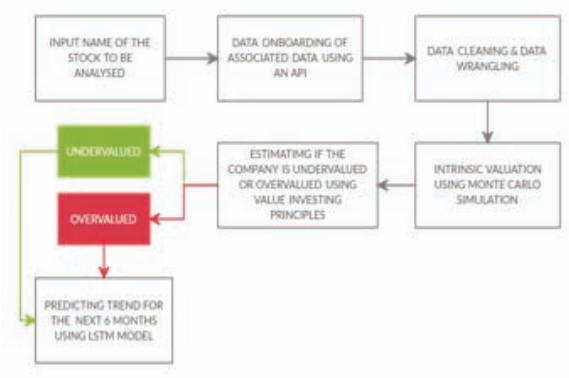


Fig. 1: Algorithm

A. Data Onboarding of associated data using an API

The API is an acronym for Application Programming Inter- face, a software intermediary that helps two systems to connect with each other. When you use an app on your cell phone , for example, the app connects to the Internet and sends the data to a server. Then the server gets the information, interprets it, performs the actions needed and sends it back to your phone. The app then interprets the information and provides you with the details you want in a readable way. All this happens through an API.

1) **Need for an API:** : It is needed for market analysis in real-time. Standard analysis methods should not be underesti- mated – using a variety of sources allows for a rich data set. However, there are downsides to traditional approaches.
- Data becomes outdated very fast as all the insights are based on historical inputs
- Traditional market analysis, especially when analysing massive data sets, is time-consuming.
- Flexibility does not account for these research methods. This is where research into real-time truly comes into its own.

2) **API used:** : Currently the proposed paper uses the pop- ular and most successful API called Alpha Vantage to fetch real-time data of any company. This data is grouped into 4 categories:-

1) Stock Time Series Data (from the last 20 years).
2) Physical/Digital Cryptocurrencies
3) Technical Indicator 4)Sector Performances

Although the proposed paper doesn't uses all the functional- ities, this API will allow us to get the Company Overview which has many financial ratios which will be very helpful for us to calculate the intrinsic value of a particular stock.

B. Data Preprocessing

1) Convert balance sheet of the company of the last 5 financial years into dataframes.
2) Then financial ratios derived from the balance sheet and put it into a data-frame with some of the important parameters highlighted. All the parameters to be con- sidered have been discussed in detail.

C. Monte Carlo Simulation for Stock Valuation

The system's key requirement is to upload the excel/CSV file consisting of many financial input parameters which are automatically uploaded. It will then generate parameters to achieve the

intrinsic value of the stock. Free cash flow to firm (FCFF) reflects the volume of cash flow from activities available to common shareholders taking into account depreciation, working capital, tax and investment expenditures. Free cash flow to the firm is an indicator of the performance of a business after all spending and reinvestment. It is just one of several measures used to compare and evaluate the financial health of a business.

Fig. 2: Valuation Block Diagram

Cost of equity (Ke) is the rate of return on that the company must pay to its investors, i.e. shareholders to cover the expense of spending its equity. Rf is the standard risk free rate for 10 years government bond yield. Return on market (Rm) of the stock calculates the profit return on the market capitalization of the firm, which is a function of the stock value and the number of shareholders. Beta (β) is a measure of portfolio's volatility, or systemic risk relative to the market. Based on its variance and market returns, it measures the potential return of an asset. Beta is called beta coefficient as well. In addition to the depreciation and subtraction of capital spending, the income after tax (PAT) is measured. Depreciation addition, that is, non-cash expenses and capital expenditure being subtracted from the statement of income of the company. Cost of Debt (Kd) is the effective interest rate that the company is paying on its current debt. Usually, it is company's cost of debt before taxes.

The weighted average cost of capital (WACC) is the value the company is expected to pay to all its existing shareholders on average to fund its assets. WACC is commonly referred to as the capital cost of the organisation we are looking at. Importantly, it is determined by the open market, not by administration. The weighted average capital cost reflects the expected returns that a company must receive on an established asset so that they can sustain its creditors, investors and other capital providers.

Terminal growth is the perpetual price at which businesses are projected to increase free cash flow. The disparity in inflation rates is between 2% and 3%. It is determined using Gordon Growth model. The present value for the shareholder of any potential cash flow forecasts for which the stock is expected to be produced over time is determined. In turn, free cash flow to shareholders and fair market value to equity owners is given by summing up the free cash flow. The present value per share shall be determined by dividing the present value of the shareholder by the number of outstanding shares. The present value per share is determined on the basis of the financial statements of the last five years. Now, the Monte Carlo simulation is used to estimate the intrinsic value of the company. For this purpose, the current market price will be used and and the range is selected for the present value per share and the market value. The mean and standard deviation of the shift in range values is calculated for simulation purposes. After the simulations are made, the final value of the company is determined and, as a result, the performance of the company is evaluated.

D. Market Trend prediction using LSTM

In sequence prediction cases, LSTMs (Long Short Term Memory) are really effective because they can be used to preserve past information. In this paper, this is essential because the previous price of a stock is key to its future market prediction.

The stock market data will be fetched from Yahoo Finance. This website provides URL-based APIs that provide historical stock data collected from different companies by simply specifying certain parameters in the URL.

The API fetched data will consist of 6 features: 1) Date: Date of data collected

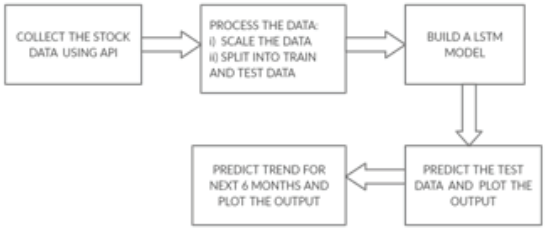


Fig. 3: Prediction Block Diagram

2) **Open:** Stock's price when market opened
3) **Close:** Stock's price when market closed
4) **High:** Highest intra-day price achieved
5) **Low:** Lowest intra-day price achieved
6) **Volume:** Number of shares transacted in the market throughout the day

The above data needs to be transformed into a suitable format for use with prediction model by scaling the data to the [-1, +1] range. The input dataset is split into training and test datasets.

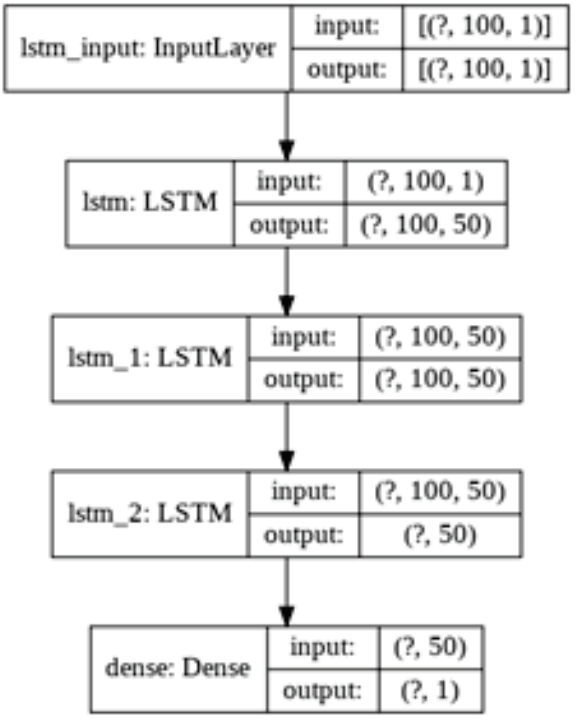


Fig. 4: Architecture of the model

The figure above reflects the model's architecture that is used to make predictions. The LSTM is a single input layer network making five neurons 'n' hidden layers (with 'm' LSTM memory neurons per layer), and one output layer (with one neuron). One input layer, three hidden LSTM layers and one Dense layer are used for outputting the predictions. The model is fitted to the training dataset. For evaluation of the model, Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics are determined.

MAE	MSE	RMSE
0.0239	0.0011	0.0317

Performance metrics

Its performance was evaluated on two datasets. First, the train data and second, the test data were predicted. Here, the blue plot represents the market

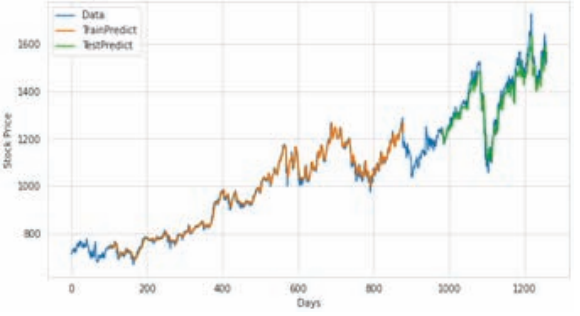


Fig. 5: Performance of the model

trend, i.e. scrapped using the API while the orange plot represents the forecast for the training dataset. The green plot is a prediction of the test dataset. The metrics and above figure shows that the model has made a decent estimate about the training and test dataset. The model can now be used to predict the company's trend for next few months.

Fig. 7: Microsoft Corporation

Similarly, the next figure indicates that the intrinsic value for Microsoft Corporation. It is 162.94 USD while the actual market price is 202.47 USD (as at 30/10/2020). This clearly illustrates us that the share of Microsoft Corporation is overvalued by USD 39.53 USD.

Results

A. Intrinsic Value

The valuation of the company was carried out after an alyzing the financial statements of the past 5 years. The intrinsic value was calculated for two US-based companies, namely Apple Inc.(T icker : AAPL) and Microsoft Corporation(T icker : MSFT), using the DCF model and Monte Carlo Simulation. In order to make a long-term de-cision, compare the estimated intrinsic value, also known as Fair Price, with its current market price.

B. Market Trend and Intrinsic Value

Here, the market dynamics for Apple Inc. and Microsoft Corporation over the past five years have been visualised along with their intrinsic value line (red line). This illustrates, ultimately, how the company's share has been carried out over the years.

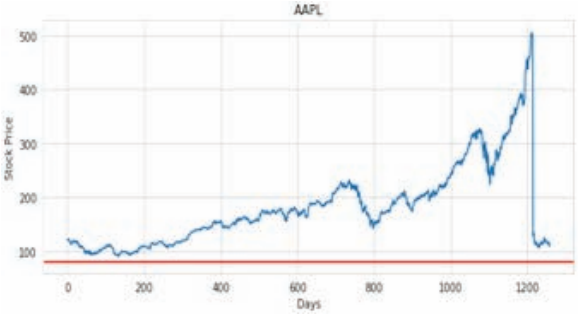


Fig. 8: Market Trend for Apple Inc.

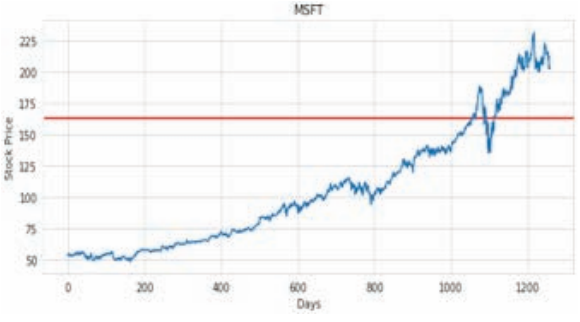


Fig. 9: Market Trend for Microsoft Corporation

C. Market Trend Prediction using Stacked LSTM

The model is ready to make future predictions after the testing phase. It is used to generate future predictions for the next 6 months.

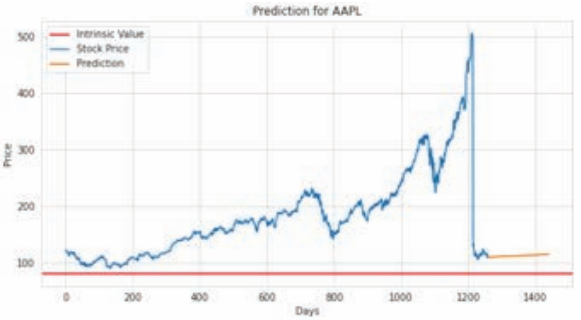


Fig. 10: Prediction for Apple Inc.

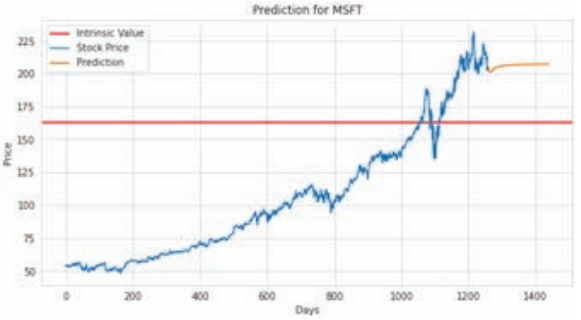


Fig. 11: Prediction for Microsoft Corporation

C. Market Trend Prediction using Stacked LSTM

The model is ready to make future predictions after the testing phase. It is used to generate future predictions for the next 6 months.

The figures above are of the predicted trend developed as model's output.

Conclusion

The results obtained clearly indicate that together with Monte Carlo simulation, the use of DCF models has proved to be a boon for long-term investment research. Although this method can be sufficiently scaled to apply to enterprise computing systems, the proposed paper also takes into account the current market rate which is unprecedented. The results show that this novel method of calculating intrinsic value lies in the ballpark of various existing approaches. Furthermore using LSTM Neural Networks, an investor using this tool, goes a step

ahead of other investors in terms of knowledge about the company. This prediction will help the investor to make a wise investment decision.

It should be noted that the proposed system has scope of improvement, particularly in prediction of the company's stock price over a period of time. The current proposed system predicts the stock price up to only 6 months accurately beyond which the model under-performs its current state. Accurately predicting the time of 'cross-point' i.e. the time (date) at which the stock price will cross the intrinsic value line is surely an area for further research. The aforementioned can be noted as the limitations of this proposed paper.

References

[1] *Coelho, Joseph, et al. "Social Media and Forecasting Stock Price Change." 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). Vol. 2. IEEE, 2019.*

[2] *S. Sangsavate, S. Tanthanongsakkun and S. Sinthupinyo, "Stock Market Sentiment Classification from FinTech News," 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE), Bangkok, Thailand, 2019, pp. 1-4, doi: 10.1109/ICTKE47035.2019.8966841.*

[3] *S. S. Siddiqui and V. A. Patil, "Proposed system for estimating intrinsic value of stock using Monte Carlo simulation," 2017 International Confer- ence on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 723-729, doi: 10.1109/ICCONS.2017.8250558.*

[4] *E. Rodrigues Reis and J. Sim a˜o Sichman, "MAVIS: A Multiagent Value Investing System," 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sa˜o Paulo, 2018, pp. 372-377, doi: 10.1109/BRACIS.2018.00071.*

[5] *L. Gao, D. Kampas and K. Rinne, "Capturing Investor Sentiment: Advancing Predictability in Finance with Computer Science Approaches," 2018 IEEE 20th Conference on Business Informatics (CBI), Vienna, 2018, pp. 97-99, doi: 10.1109/CBI.2018.10052*

[6] *I. Kumar, K. Dogra, C. Utreja and P. Yadav, "A Comparative Study of Supervised Machine Learning Algorithms for Stock Market Trend Predic- tion," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 1003- 1007, doi: 10.1109/ICICCT.2018.8473214.*

[7] *S. Kao-Yi, Y. Min-Ren and C. Kai, "A fuzzy-MCDM based value invest- ing method for banking stocks evaluation," 2010 2nd IEEE International Conference on Information and Financial Engineering, Chongqing, 2010, pp. 161-165, doi: 10.1109/ICIFE.2010.5609275.*

[8] *Schwager, Jack D. Market wizards, updated: Interviews with top traders. John Wiley Sons, 2012.*

[9] *Graham, Benjamin, and Bill McGowan. The intelligent investor. Harper Collins, 2005.*

[10] *Lynch, Peter S., Peter Lynch, and John Rothchild. One up on Wall Street: How to use what you already know to make money in the market. Simon and Schuster, 2000.*

The Future of Supply Chain – Data Logging Via Internet of Things (IoT)

Savita Singh*
Abhijeet Kumar Singh**

ABSTRACT

In the era of 1990, the computers connected via internet were very less around 30,000 to 40,000 but as the time gets changed the numbers gets increased to around 3 million in the year of 2000, and after a decade in 2010, the digits of million gets converted to 2 billion around the year of 2010, and after some time in 2016, around 2 out of 10 people is capable and using the smart home appliances like automatic washing machines, water purifiers, air conditioner, Smart bed, Smartwatches, and many more home Appliances. In 2017, the people become habituated to iPhone's "Siri". There are many more such applications that are changing our lives.

The future of entire world is only Internet of Things (IoT), that will transform the real-world objects to an intelligent virtual objects like after a decade mainly in the period of 2030 to 2040, your smart appliances may inform you about what steps to be taken at a required time and it may also inform you about what is need of them. IoT is mainly a system of web associated objects ready to gather and exchange data or we can also say that internet of things (IoT) is a worldwide system foundation with self-designing capacities dependent on principlesand interoperable communications protocols.

This paper mainly focusses on the future of supply chain – data logging via the internet of things (IoT). Data logging is the process of data collection and storage over a while to analyze specific trends or to records the data based events/actions of a system, network or IoT environment. And apart from this, the data logging also allows information security (is) and auditing staff to analyze system access information and identify suspicious activities.

The uses of data loggers are required for a multiple of reasons, frequently to ensure compliance with industry-specific regulations, and quality and environmental control procedures. As the time is changing and technology is also changing day - by – day and due to this the data logging is influencing the supply chain in the form of warehouse, logistics trucks delivering the products, to supermarkets chillers in retail environments, new cutting edge monitoring solutions on the market provide opportunities that incredibly streamline the procedure, and many more.

As a result, an enormous amount of data being generated, stored and being prepared into helpful activities that can "command and control" the things to make our lives with a lot simpler and more secure. For solving up these problems the methods used like the makings of projects, using a mix of public and private infrastructure also can help protect data, using an independent infrastructure such as cellular service to send data, LTE-M and LTE-NB use existing cellular towers and it provides much broader coverage.

**Assistant Professor, IPEM Group of Institution, Ghaziabad, sheoran.savita@gmail.com*
***Student, IMS Ghaziabad, abhijeet.singh@imsuc.ac.in*

The IoT based applications have huge advantages and will help people and various organizations in makings of smart homes, smart cities, efficient use of electricity and energy, better management and healthcare, wearables, connected car, security, road safety, cost-efficient business operations, etc.

This all is "The Internet of Things" applications and this is the Future.

Keywords: *Internet of Things, Supply Chain, Data Logger*

Internet of Things (IoT)

Is powered by the combination of analytics, mobile computing and cloud services. Asset tracking is also influenced by IoT which is providing such gadgets to make better decisions, saving money and time. Asset tracking, which gives them the tools to make better decisions and save time and money. Newer asset tracking solutions replaced the traditional bar code scanner with radio frequency identification (RFID) which provides more essential and usable information when paired with other IoT technologies.

Warehouses are also not untouched by the capabilities of IoT where inventory monitoring, stock level distribution are taking help from new models.

Networks powered IoT technologies are useful in minimizing the human error when it comes to inbound and outbound packages as it provides several scales to scan different parameter like size, weight, density, etc. based on this collected data, accuracy can be ensured by shipping collaboration software with expected and actual received inventory.

IoT Key to Efficiency

- **Operational efficiency:** Deficiencies that occurred at a real-time can be traced quickly or rectified. Companies are affected by slowdown or delay that cost them money can be identified.
- **Inventory management:** IoT Devices provides automation to the organizations for inventory updates that when to reorder or restocked. This eliminates the delay and provides the product as per the customer's expectation on time.
- **Customer service:** IoT devices are reducing the time amount from requests to respond. On time data accessibility is required to match up with the customer's expectations also the accurate delivery date time notifications are demanded by the customer as to where their product during the transit.
- **Loss management:** IoT devices make use of sensors that can sense almost all possible loss events that can occur with the product during the transit and if any such event occurs with the product that can be traced that at what point of time, place and factor that has contributed to merchandise loss.
- **Visibility:** A much better understanding of the product that exactly what amount of time and place product is stayed, equipped Supply chain management professionals with the data they require to make better decisions. Earlier, companies would get only occasional updates and outdated reports and it would be too late to make any real changes or adjustments.

Examples of IoT Transforming Supply Chain

- **New Jersey Transport Authority (NJTA)** – Undoubtedly IoT provides tools for cost-cutting in the long run. The NJTA is working with IBM to deploy 3,000 sensors along the New Jersey. Turnpike – one of the busiest roadways in the U.S. The data this produces is utilized by the crisis administrations and traffic management operators so they can get to a mishap quicker, and decrease congestion development.
- **Amazon** – Amazon handles their orders uniquely regardless of their shapes or size with the help of robots and AI system; they use Wi-Fi-connected robots to identify products by reading QR codes from built-in cameras. Workers of Amazon works in coordination with the robots and IoT devices. The priority of the product is identified by the AI systems, humans

perform restocking and packaging while the rest of the work is performed by the robots.

- **Volvo** – For shipping the ordering components to vehicles across the globe Volvo utilizes the cloud services with IoT technology to enhance its logistics of the supply chain. The company established a relationship with Microsoft, which involved trialing its mixed reality headset, HoloLens. It is the company that believes that their headset can help in transforming car design with a better relationship with the customer.

IoT and Big Data

The number of internet users and connected devices are increasing rapidly and influencing our daily life to incorporate with the Internet of Things (IoT) and Big Data. IoT devices share a huge amount of data as they are the Physical devices that are connected to the internet.

According to a study by Gartner, the revenue generated from IoT-enabled services and products will exceed $300 billion by 2020.

For any organization, well-analyzed data is at most priority and IoT is generating a huge amount of data. To analyze that hugely generated data from IoT devices Big Data analysis and analytical tools are required. Big Data analytics tools help to generate and store the insight from the information received from various sensors of IoT devices.

Predictive analytics is possible with several machine learning algorithms that use the patterns and trends observed from the vast amount of data generated by IoT sensors. Predicting the problem before it happens, so that it can be fixed is possible with Big Data Analytics. The risk of damage and waste can be minimized with Big Data leads. Thus the IoT services with Big Data creates ample opportunities to enhance customer relationship.

The following statistics from Gartner shows how IoT and big data is revolutionizing our everyday lives:

- By 2020 every person will create 1.7 MB of data every single second.
- There are 3.5 billion Google searches per day and 400 hours of new YouTube video added every minute.
- The number of IoT connected devices is forecast to reach up to 30 billion by 2020.
- IoT investment is expected to reach $58.14 trillion in the next 15 years.

IoT & Big Data in the Supply Chain

According to a report from Transparency Market Research, the global supply chain and logistics market are set to exceed $15 trillion by 2023.

Although there is rapid growth in the supply chain industry, still there are not that many field innovations happening and the companies still lacking efficiencies.

Based on a report by Zen Cargo , supply chain inefficiencies cost businesses nearly USD 2 Billion in the UK alone.

The traditional outdated process in the supply chain is very complex involving controlling and monitoring product flow row material to final product delivery with a point to point communication that relies on e-mail and phone communication. While the controlling and monitoring of the product is very crucial. This can cause inefficiencies as the speed of supply chain is slow-down by the big network of point-to-point communications.

"IoT is on the rise towards restructuring the entire process by which supply chains operate."

A smart network ecosystem of people, process and data through sensors and actuators, that is consistently collecting, measuring and distributing real data, is the power of the Internet of things. This real data gives its benefits to the supply chain providing visibility in every process within the supply chain.

Why the Internet of Things Matters to the Supply Chain?

IoT devices are a major advantage in aspects of supply chain management:

- Visibility and tracking of real-time shipment and inventory
- Stakeholders can easily plan supply and demand as they know when they can expect to receive and process goods

- Early identification of issues with lost or delayed goods
- Keeping raw materials and processed goods in optimal conditions provides enhanced quality management.
- Assurance of goods location in rest or motion as per stakeholder.
- Better storage and distribution of products.

IoT Enabled Data Loggers in Supply Chain

Technology is advancing at a rapid rate and companies still working with obsolete technologies are holding back. The same scenario is prevalent in supply chain industries. Traditionally, data loggers were having significant drawbacks although they were widely used in supply chains and covers basic cargo monitoring needs.

The old traditional data logging technology widely used in the supply chain with electronics devices (data loggers) used to log location-based environmental data was considered as state of the art innovative technology of the time but with the time that was affected with new trends and became obsolete. Traditional data loggers were replaced by IoT enabled solutions providing more vital real data with wider visibility.

IoT enabled devices vs. traditional data logging technology…

1) Intelligent vs Non-Intelligent Analytics

The key component of IoT enabled data logging technology is the analytics, providing information on shipment, help in making decisions, future predictions and exposing risks. It can generate intelligent reports based on performance management with various quality checks in less time, such technology empowers the business with better decision making ability. On the counter side, the traditional data logging technology using few independent electronic devices is not that intelligent that it can generate such future prediction reports on its own. Such devices are used for collection and data storage, the older data logging devises are not bothered about the predictions on shipment and related risk management neither with the financial management.

2) Instantaneous Data Exchange + Analytics vs Limited Data Accessibility

The power of IoT enabled devices is in its sensing, analytics and communicating the real data with all of its related stakeholders instantaneously. The IoT enabled devices to make use of a cloud based dashboard and provides the information to all related parties in real-time. Whenever cloud dashboard updates, all parties get to know with the recent information, which makes easy communication between them. On the other hand, old data loggers are not capable with such capability and communication technology of sending real-time data to all. Here data is stored in one device which along with the shipment. That data can be exchanged only when it is extracted from the device.

3) Automatic Data Transfer vs. Manual Setup

The IoT enabled technology transfers the data on the go which enables the problem-free quality condition monitoring of the product. Without any delay, the product reaches the destination it will be delivered to the buyer. Here no other means of IT infrastructure is required for data transfer and analysis as all the process is completed by the means of automation over the cloud. Whereas data extraction and analysis are not like that much easy in traditional data logging technologies since here dedicated software with data wiring is required to be installed at the destination to extract and analyze the data, sometimes this cause problem in case of quality condition monitoring as the extra time needed for data extraction and analysis.

4) Real-Time Data Stream Vs. Data Availability Post Shipment

Any disruption with the product can lead to affect customer satisfaction and business operation, IoT enabled devices provides all real-time updates related to the shipment and enables you to react against such disruptions. The traditional data logger technology is unable to act with real-time data as it provides the details only after the shipment reaches the destination. Traditional data loggers are not capable to reveal important conditional measures sometimes which are of utmost priority.

The IoT Transport Data Logger

TDL or the Transport Data Logger gives transparency in the supply chain process. TDL is

moved with the shipment and measures several parameters like temperature, shock, and tilt. Different measures on those parameters can be settled and if any of the parameter is breached then that will be traced in the supply chain. TDL also facilitates data visualization through mobile applications.

Benefits of the Transport Data Logger

- Efficiency: Easy to use and configure, easy integration without the pre-requisite of the logistic chain.
- Condition Monitoring: 360-degree approach of TDL with condition monitoring makes it more transparency in the supply chain. When any parameter threshold is exceeded, the TDL acquaints with verifiable proof of possible primary and secondary damage.
- Simplicity: IT is a reliable, versatile, simpleand cost-effective delivery monitoring device.
- Transparency: The TDL creates trust between parties and provides data for enhancing the logistics process. It offers proof of a fail-safe transport chain.

Future of supply chain with IoT

IoT today: an information flood that carries constant reaction to changing client needs and economic situations with late gauges of 28 billion IoT associated gadgets worldwide by 2021 the main thing IoT will do is add to the blast of data driving organizations' information intricacy challenge. Simultaneously, IoT will challenge supply chains to open up to a new plan of action and operational potential outcomes. These are empowered by IoT information streaming once more from clients as an immediate contribution from organized sensors joined to conveyed items, just as from a huge number of outside sellers.A foundation of this vision is that prescient investigation will caution organizations to issues rising with their gadgets in client use, and afterward, important inventory network procedures can be marshaled to react to the client — potentially even before the client becomes mindful of the issue. Supply chains reacting to changing client needs progressively adequately change items into "items as-an administration" — another advanced plan of action.

IoT, tomorrow: wise, self-arranging supply chains It is a little reasonable jump from items as-a support of shrewd, self-sorting out supply chains. As production network forms and their crude materials and segments become instrumented with IoT sensors, the sign they send about the condition of those procedures can be investigated by progressively proficient AI frameworks. Joining that information with data about the different clients for whom the store network's yield is predetermined, such frameworks could choose for themselves how to work and react progressively to evolving conditions.

References

[1] *:https://www.govtech.com/fs/New-Jersey-Turns-to-Internet-of-Things-to-Improve-Roadway Safety.html*

[2] *:https://www.govtech.com/fs/New-Jersey-Turns-to-Internet-of-Things-to-Improve-Roadway Safety.html*

[3] *:http://www.informationweek.com/strategic-cio/amazon-robotics-iot-in-the-warehouse/d/did/1322366*

[4] *:https://www.media.volvocars.com/global/en-gb/media/pressreleases/169675/volvo-cars-todevelop-next-generation-automotive-technologies-with-microsoft*

[5] *:https://www.gartner.com/en/newsroom/press-releases/2017-02-07-gartner-says-8-billionconnected-things-will-be-in-use-in-2017-up-31-percent-from-2016*

[6] *:https://www.google.com/search?q=gartner+iot+study*

[7] *:https://www.prnewswire.com/news-releases/global-logistics-market-to-reach-us155-trillionby-2023-research-report-published-by-transparency-market-research-597595561.html*

[8] *:https://www.google.com/search?q=ZenCargo+study+supply+chain+inefficiencies+cost+businesses+nearly+USD+2+Billion+in+the+UK+alone*

[9] *:"IoT Will Surpass Mobile Phones as Most Connected Devices," InformationWeek, 4 August 2016, © 2016 UBM.*

[10] *;Sheng, Z., Yang, S., Yu, Y., Vasilakos, A., Mccann, J., & Leung, K. (2013). A survey on the ietfprotocol suite for the internet of things: Standards, challenges, and opportunities. IEEE WirelessCommunications, 20(6), 91-98.*

[11] Theoleyre, F., & Pang, A. C. (Eds.). (2013). Internet of Things and M2M Communications. RiverPublishers.

[12] Coetzee, L., &Eksteen, J. (2011, May). The Internet of Things-promise for the future? Anintroduction. In IST-Africa Conference Proceedings, 2011 (pp. 1-9). IEEE. International Journal of Computer Science & Information Technology (IJCSIT) Vol 10, No 2, April 2018.

[13] Ji, Z., &Anwen, Q. (2010, November). The application of internet of things (IOT) in emergencymanagement system in China. In Technologies for Homeland Security (HST), 2010 IEEEInternational Conference on (pp. 139-142).

AI Assistant to support students and the users

Nandini Bagga*
Pratikshit Vashistha**
Palak Yadav***
Dr. Arvind Kumar****

ABSTRACT

Artificial Intelligence(AI) Virtual Assistant is an application program that understands natural language voice commands and completes tasks for the user. Our project is based on an AI Virtual Assistant for Galgotias University. This Virtual Assistant will tell us about the information regarding the curriculum and academics of Galgotias University including all the courses at Galgotias university and its Student Club Organizations. It will use speech-to-text, so that the program understands our command in text form and it will give us the result by text-to-speech. There is a great role of python libraries in this, as it allows us to add a speaker to our program which answers us in speech form.

Keywords—*Intelligent Personal Assistant, Automation, Python libraries, Personalized, Galgotias University*

Introduction:

The utilization of an Intelligent Personal Assistant to improve learning is an arising practice that, despite the fact that not yet far reaching, has a significant future job. The usage of IPAs through Voice User Interfaces implies that these partners can give prompt and instinctive reactions to normal language boosts, so the client can create voice collaboration through the PC framework. Furthermore, a considerable lot of them incorporate the chance of making applications at no expense for their turn of events and use, for example, the Amazon Echo or the Google Home partners.

Overview

A. Main Issue

The main issue that we wanted to deal with was to simplify the whole process of performing basic operations and to make it convenient for different classes of computer owners; the tech savvy users, the busy users, and even disabled users. To be able to access information to automate Galgotias University Applications using voice commands. Students can check their marks, attendance,etc. Same applies for teachers. The idea of a VPA becomes attractive as it changes the focus of the supporting system to the contextual sphere under private control of the user.

B. Our Galgotias University AI Assistant

We want to create something that saves time and so that we don't have to spend extra time on just searching about the details of Galgotias University. In order to Perform more complex tasks than others in fraction of seconds using artificial intelligence technology. People can easily access all the information regarding courses of the branches, to events and much more, happening in the university.

- A great source for just-in-time information.
- Ensures accuracy and does work faster.
- With AI, we can learn from the user; we can understand what they're good at, what they struggle with, and adapt. This allows us to show them only information that is relevant to them, personalised.

*B.Tech CSE- AI & ML, Galgotias University, Greater Noida, India bagganandini8@gmail.com
**B.Tech CSE- AI & ML, Galgotias University Greater Noida, India, pratikshitvashistha@gmail.com
***B.Tech CSE, Galgotias University Greater Noida, India, palak.yadav2224@gmail.com
****Associate Professor, Galgotias University Greater Noida, India, arvindkumar@galgotiasuniversity.edu.in

- We can collect data and analyse at broad level in no time which would take days doing manually.
- Can instantly add, update or remove.
- We'll build assistant so well that the learners would be able to have conversations with them casually and have solutions without travelling from one block to another, hence leading to zero wastage of time.
- It'll definitely avoid the leak of private and sensitive information regarding staff and students.
- We can provide our specific campus driven environment for its working.

Explanation

In the previous sections we have identified the issues and the roles of an intelligent assistant to support students and other users, and to help them to overcome difficulties while using several university applications and how you can easily get intel about the University..

A. Architecture of Our Assistant

The architecture of our assistant has four main modules.

1) **User:** This module manages all the information about the user, his environment. Each user can be defined by the system that they use, each system is defined as the user here.
2) **Conversation Agent:** This module is the human-computer interface; the interaction with the user is in natural language.
3) **Web/ Internet:** The web module is required to make the searches and it manages all the web/ Internet related queries.
4) **GU ICLOUD and LMS:** GU Icloud and LMS are the two most frequently used applications offered to the students by Galgotias University. This module manages the web-automation of these applications.

- **User Module**

The goal of the user's profile is to store a student's details. The management of user description is essential to an adaptive and personalized system.

The user uses this program for their personal use which can be web searching, playing songs, taking notes, and if the user is a student or a teacher that belongs to the Galgotias university will be able to access the information like attendance, results and other stuff of the GU Iclouds site and LMS site. The application or the program should be installed or should run in the system of the user to access the advantages of the application.

- **Conversation Agent**

The interaction between the user and assistant is carried out by a virtual agent.

Its objectives are:

1) To manage the dialog model to communicate with the user.
2) To interact with the user, take the command verbally by the user and give the desired answer in speech format.
3) Take different commands from the user and perform the task, like for gathering intel or for opening an application or any other operation of automation.

- **Web/ Internet Module**

This module is based on the API and JSON file which is used by the library of the python program. The web is required to make the searches and it is highly used because of the reorganization of the voice and the sentence or the words used by the user and act according to the voice input given by the user to the program.

It is used for automation of processes associated with the applications used by the University, to make the process of viewing your attendance, results, marks, etc., easier, simpler and less time consuming.

- **GU ICLOUD and LMS**

This module is applicable for the user that belongs to the Galgotias University which will be defined as :

a. Student :

Each student belongs to a class identified by semester and section. Each class belongs to a department and is assigned a set of courses. Therefore, these courses are common to all students of that class. The students are given a unique username and password to login. Each of them will have a different view. These views are described below.

- Student information - Each student can view only their own personal information. This includes their personal details like name, phone no., address etc. Also, they can view the courses they are enrolled in and the attendance, marks of each of those.
- Attendance information - Attendance for each course will be displayed. This includes the number of attended classes and the attendance percentage.
- Marks information - There are events and semester end examinations for each course. The marks for each of these will be provided in the system by the assistant.
- Notifications and events - This section is common to all students. Notifications are messages from the admin such as declaration of holidays, test time-table etc.

b. Teachers:

Each teacher belongs to a department and are assigned to classes with a course. Teachers also have a username and password to login. The different views for teachers are described below.

Information - The teachers will have access to information regarding the courses and classes they are assigned to. Details of the courses include the credits, the syllabus plan. Details of the class include the department, semester, section and the list of students in each class. The teacher also has access to information of students who belong to the same class as the teacher.

Attendance - The teacher has the ability to add and also edit the attendance of each student. They can enter the attendance of the whole class on a day to day basis. Teachers can edit the attendance of each student either for each student individually or for the whole class.

B. Technologies

- **Artificial Neural Networks for speech recognition -**

Neural Networks form the base of Deep Learning. Algorithms of Neural Networks are based on the structure of a human brain. Hidden Layers perform most of the computation required by the network. Each neuron is connected to the neurons of the next layer through channels. Each channel is assigned a weight. Activated neurons transmit data to the neurons of the next layer. Wrong predictions are figured out by the error. Weights are adjusted during back propagation. Combination of Back propagation and front propagation is applied to make a right prediction.

Feature Extractor - If a noisy signal is received, the objective is to produce a signal with less noise; if image segmentation is required, the objective is to produce the original image plus border maps and texture zones; if lots of clusters in a high dimensional space must be classified, the objective is to transform that space such that classifying becomes easier, etc.
Recognizers deal with speech variability and account for learning the relationship between specific utterances and the corresponding word or words. There are several types of classifiers but only two are mentioned: the Template approach and the approach that employs Hidden Markov Models.

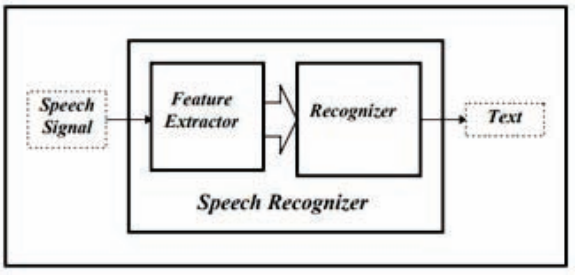


Figure 1: Basic Model of Speech Recognizer

- **Python Libraries -**

- **pyttsx3:** pyttsx3 is a text-to-speech conversion library in Python. Unlike alternative libraries, it works offline, and is compatible with both Python 2 and 3.An application invokes the pyttsx3.init() factory function to get a reference to a pyttsx3. The pyttsx3 module supports two voices: first is female and the second is male which is provided by "sapi5" for windows. sapi5 – SAPI5 on Windows.
- **Speech recognition:** This package can be installed by using pip install Speech Recognition.To use all of the functionality of the library, you should have: Python 2.6, 2.7, or 3.3+ (required), PyAudio 0.2.11+ (required only if you need to use microphone input, Microphone).
- **os:** This module provides a portable way of using operating system dependent

functionality. If you just want to read or write a file see open(), if you want to manipulate paths, see the os.path module, and if you want to open a file use startfile(path).

- **smtplib:** The smtplib module defines an SMTP client session object that can be used to send mail to any Internet machine with an SMTP or ESMTP listener daemon. We used smtp.login() to login to the user's account and sendmail() to send email from the user's account. smtp..ehlo() Identify yourself to an ESMTP server using EHLO.
- **web browser:** The webbrowser module provides a high-level interface to allow displaying Web-based documents to users. Under most circumstances, simply calling the open() function from this module will do the right thing.
- **wikipedia:** Wikipedia is a Python library that makes it easy to access and parse data from Wikipedia.
- **Psutil:** this library is used to get the access for the cpu status.
- **Pyjokes :** this library contains jokes and this makes the virtual ai assistant to speak out different jokes when this function is called
- **Json:** this is used to fetch the data from the api.
- **Selenium:** Automates Web browsers. Selenium WebDriver is a collection of open source APIs which are used to automate the testing of a web application. Description: Selenium WebDriver tool is used to automate web application testing to verify that it works as expected. It supports many browsers such as Firefox, Chrome, IE, and Safari.
- **ChromeDriver:** WebDriver is an open source tool for automated testing of web applications across many browsers. It provides capabilities for navigating to web pages, user input, JavaScript execution, and more. ChromeDriver is a standalone server that implements the W3C WebDriver standard.

Related Work

There is a substantial body of work investigating the use of Artificial Intelligent assistant to help users perform various tedious operations with their voice assistants. Much of the work has focused on creating assistants surrounding the topic of a university, topics including students, curriculum, studies, etc., while ignoring to make an assistant for the whole university itself, regarding the intel and the application systems used by the university.

SwiftFile's philosophical underpinnings are probably most related to those of CAP (Mitchell et al. 1994), calendar-scheduling application. CAP learns to predict how users will respond to the various questions it must ask in scheduling a meeting and offers its best predic- tion as a default value. If CAP's predictions are correct, the user can simply hit return to accept the suggested value. Otherwise, the user may override the default by typing in a different response. Both SwiftFile and CAP are unobtrusive assistants because they offer con- venient, overridable shortcuts to the user rather than taking possibly incorrect actions on the user's behalf. Payne and Edwards (Payne and Edwards 1997) do consider the possibility of incremental learning, but very few details are provided.

Future Work

Future work includes, to use machine learning algorithms, so that the assistant can be more efficient and so that it can come up with answers, asked by the user,on its own. Reinforcement machine learning models can be used to make the assistant come up with its own reasoning. Reasoning model can be made which will have fuzzy-logic based inference mechanisms, and machine learning tools are added to detect when a student encounters difficulties. The early identification of failure is the key success factor of our system. The reasoning module contains the student's diagnosis and the non pedagogical helps modelling. The Basis model which will maintain personal information as interests, preferences; cognitive profile, learning style and schooling's data: identification, personal data, interest, preferences learning results. The Environment model keeps information about the student's work context: type of device, place, and time. The Information System (IS) model handles the information about the technical environment in which the assistant is integrated, for example a Virtual Learning Environment. And the information about the student's learning, for example the school's organisation.Our next step is to make these

modules work together to build a new prototype of our non pedagogical agent, test and validate it in real-world applications.

Final Remarks

Galgotias University(GU) AI Assistant is an easy-to-use personal assistant that helps users to get all the information related to the university and manage their university applications. GU Assistant makes very few demands on users. GU Assistant has virtually no adverse side effects. We intend to include the future works as mentioned above. GU Assistant can be considered as a companion that simplifies the processes involved with the admission procedures, information related to GU and GU applications

References

[1] *AN INTELLIGENT ASSISTANT TO SUPPORT STUDENTS AND TO PREVENT THEM FROM DROPOUT. Tri Duc Tran, Bernadette Bouchon-Meunier. Christophe Marsala and Georges-Marie Putois, LIP6 DAPA, Université Pierre et Marie Curie, 104 Avenue du Président Kennedy, Paris, 75016, France*

[2] *SwiftFile: An Intelligent Assistant for Organizing E-Mail. Richard B. Segal and Jeffrey O. Kephart, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598*

[3] *Stack Overflow - Where Developers Learn, Share, & Build Careers*

[4] *Python | Text to Speech by using pyttsx3 - GeeksforGeeks*

[5] *Build A (Full-Featured) Instagram Bot With Python - YouTube*

[6] *An Intelligent Assistant for Patient Health Care Silvia Miksch, Kenneth Cheng, Barbara Hayes-Roth*

[7] *NOEMON: An Intelligent Assistant for Classifier Selection*

[8] *An Intelligent Assistant for Computer-Aided Design. Extended Abstract. Olivier St-Cyr, Yves Lespérance, and Wolfgang Stuerzlinger*

[9] *FRM: An Intelligent Assistant for Financial Resource Management.*

[10] *An Intelligent Assistant for the Architectural Design Studio.*

[11] *Surgery task model for intelligent interaction between surgeon and laparoscopic assistant robot.*

[12] *Expert System as an Intelligent Assistant for Computer Users*

[13] *Design Considerations for a CBR-based Intelligent Data Mining Assistant.*

[14] *INTELLIGENT CURRICULUM ASSISTANT. Conference: Proceedings of 11th International Conference on Education and New Learning TechnologiesAt: Palma, Mallorca, Spain*

FEEDBACK FORM

Your valuable comments will help us to shape the future issues better

	Highly Appreciable	Somewhat Appreciable	Not Appreciable	Did Not Read
Application of Machine Learning for Predictive Analytics:Indian Premier League (IPL) T-20 Cricket Matches	☐	☐	☐	☐
Power System Security Assessment using K-Nnearest Neighbour	☐	☐	☐	☐
Sementic Information Retrieval:Unboxing the Complexity in Big Data	☐	☐	☐	☐
Social Media Analytics Platform to assist Business Decision Making for Small and Medium Enterprise in Indonesia	☐	☐	☐	☐
A review paper on Identification of the CAPTCHA with the advancement of Machine Learning Techniques	☐	☐	☐	☐
Customer Behaviour Prediction using Propensity Model	☐	☐	☐	☐
Estimate of Modeling Units for Hindi Speech Recognition using Artificial Intelligence	☐	☐	☐	☐
Crypto currency: A Futuristic Digital Currency for the Digital World	☐	☐	☐	☐
Blockchain with Artificial Intelligence	☐	☐	☐	☐
Investment Analysis based on the Principles of Value Investing using Machine Learning	☐	☐	☐	☐
The Future of Supply Chain – Data Logging Via Internet of Things (IoT)	☐	☐	☐	☐
AI Assistant to support students and the users	☐	☐	☐	☐

Comments/Suggests(if any): _____

Name :Mr./Ms. _____

Designation: _____ **Organization/Institution:** _____

Address: _____

Phone: _____ **E-mail:** _____

SUBSCRIPTION FORM

I wish to subscribe to / renew my subscription to " IPEM Journal of Computer Application & Research" for 1 / 2 / 3 years(s).

A bank draft /cheque* bearing no…….....dated…….....drawn in favour of INSTITUTE OF PROFESSIONAL EXCELLENCE & MANAGEMENT, payable at GHAZIABAD / DELHI towards subscription for …….. years, is enclosed.

Name: ……..

Org. / Ins.: ……

Address: ……

City: ……Pin……..

Phone: ……

Fax: ……

Mobile: ……

E-mail: ……

Category: ……

Year: ……

Subscription Rates			
Category	1yr.	2yr.	3yr.
Indian (in Rs.)			
1. Institutions	300	550	800
2. Individual	200	350	550
3. IPEM Student/ Alumni	150	250	350

_____

Signature with date

The Editor

The IPEM Journal of Computer Application & Research
Institute of Professional Excellence & Management
A-13/1, S.S. G.T Road,
Industrial Area, NH-24 By Pass
Ghaziabad-201010
Tel.: 0120-4174500

Affix
Postal
Stamp

The Editor

The IPEM Journal of Computer Application & Research
Institute of Professional Excellence & Management
A-13/1, S.S. G.T Road,
Industrial Area, NH-24 By Pass
Ghaziabad-201010
Tel.: 0120-4174500

Affix
Postal
Stamp

NOTE

GUIDELINES FOR AUTHORS

All papers are subjected to a blind peer review process. Manuscripts are invited from Academicians, Scientists and Research Scholars for publication consideration. Papers are accepted for editorial consideration through email *journalit@ipemgzb.ac.in* with the understanding that they have not been published, submitted or accepted for publication elsewhere in the same form, either in the language of the paper or any other language without the consent of the Editorial Board.

Title : The title of the paper should be concise and definitive.
Length : Contributions should not exceed 5000 words including Charts, Tables and other Graphics. The order of the paper should be preceded by an Abstract (200-300word) ,Introduction, main text, tables and charts in black & white only(with caption) and figure captions, list of symbols and abbreviations (if necessary), conclusion and list of references.
Abstract : A short abstract (200-300) words should give a clear indication of the objective, scope, and results of the paper. Some Key words must be provided.
Authors Names and Affiliation: It is a journal policy that all those who have participated significantly in the technical aspects of a paper be recognized as co-authors or cited in the acknowledgements. A cover page containing the title, author's name & affiliation, mailing address including city, state, and zip code, mobile number, fax number and e-mail address.
Copyright: Each manuscript must be accompanied by a statement that it has been neither published nor submitted elsewhere for publication, in whole or in part.
Paper Acceptance: The final decision on publication is made by the Editor-in-Chief upon recommendation of Editorial Board members. Acceptance of papers for publication shall be informed through e-mail or through normal mail. Manuscripts that fail to conform to the guidelines will not be considered for publication. The author whose paper is published will receive one free copy of journal that carries the paper.
References: The references should be brought at the end of the article, and numbered in the order of their appearance in the paper. References should include full details of the name(s) of the author(s), title of the article or book, name of the journal, details of the publishers, year & month of publication including page numbers, as appropriate. References should be cited in accordance with the following example:

1 Mishra, K.M. (2002) Knowledge Management, New Delhi: Pearson Education MA: Allyn & Bacon.
2 Rowling,J.K.(2001) Harry Potter and the Socerer's Stone.London: Bloomsburg Children's.
3 Tyagi, R.M, and Malik, S.P. (2007) Job Satisfaction Working Paper No 46, Indian Institute of Travel Management, Gwalior.
5 Jacoby, W. G. (1994). Public attitudes toward government spending. American Journal of Political Science, 38(2), 336-361. Retrieved from http://www.jstor.org.

Publication Cost: Free Publications\No Page Charge.

Frequency of Publication: Annual

About the Institute

IPEM made a modest beginning in the year 1996, with few Management and Computer Application Programmes. Today the IPEM Group of Institutions are in the forefront of imparting knowledge in the field of Education, Law, Management and Information Technology. The Dept. of Computer Applications was started in 1996 with Bachelor of Computer (BCA), affiliated to the Chaudhary Charan Singh University , Meerut with 120 seats. This journal of Computer Applications students are exposed to emerging trends in the areas of information Technology by value additions through workshops, Live Project and a regular interaction with Experts from Industry . This is reflected in the performance of the students as we have 100% result with maximum 1st division . We provide best placement to the students.

The Dept. of Computer Applications is running two courses successfully: Master of Computer Application(MCA) is approved by All India Council for Technical Education(AICTE) and affiliated to Dr. A.P.J. Abdul Kalam Technical University(APJAKTU) Lucknow and Bachelor of Computer Application(BCA) is affiliated to the Chaudhary Charan Singh University, Meerut.

The other courses are running under IPEM group of Institution are Master of Business Administration(MBA) and Master of Applied Management(MAM) approved by all India Council for Technical Education (AICTE) and affiliated to DR. A.P.J. Abdul Kalam Technical University(APJAKTU) Lucknow. The Post Graduate Diploma in Management (PGDM) is approved by All India Council for Technical Education (AICTE) Govt. of India, Ministry of HRD. The Bachelor of Business Administration (BBA), Bachelor of Law(LLB)(3 years)BALLB(5 Years)approved by Bar Council of India and affiliated to the Chaudhary Charan Singh University, Meerut, Bachelor of Education(B.Ed.() and Basic Teacher Certificate(BTC) approved by National Council for Teacher Education(NCTE). Bachelor of Education(B.Ed.) is affiliated to the Cahudhary Charan Singh University, Meerut and Basic Teacher Certificate (BTC) is affiliated to the State Council of Educational Research and Training (SCERT) Lucknow.

The focus of IPEM has always been to be at the forefront of optimum utilization of IT resources and leverage the power of IT in making the learning process, informative and engaging. The students are provided with hands on experience and learning process, informative and engaging. The students are provided with hands on experience and learning with the state-of-the-art technology.

The Dept. of Computer Applications has enriched with well equipped labs in Aryabhatta Block i.e. Programming Lab for the specialization in Database, Java,.Net etc(Aryabhatta Lab-1), Internet Lab(Aryabhatta Lab-2) and for UNIX, LINUX, Android etc(Aryabhatta Lab-3). The Computer Applications of IPEM group of Insistitutions prepare the students who would be able to lead the future Industry and chase the world-wide mega trends. The Department has shined covered out for itself s commanding position with best results and placement.

Dept. of Computer Applications of IPEM Group of Institutions organizes various workshop and seminar on latest IT trends every year. Seminars often feature several speakers, each one providing information from a different angle or perspective. People who attend seminars learn new ideas and skills to help them improve their production, while those who present at seminars gain exposure for their products or services. Presenting at an academic seminar is an important part of a researcher's/Scholars life, and is an opportunity that most young researchers look forward to. A good mix of paper presentations and journal publications is important when looking to move up the academic career ladder as well.

Spacious Lecture Theaters are thoughtfully designed to induce high quality learning and are equipped with high and teaching aids such LCD and OHP projectors. Priority is attached to achieve optimal convergence of stimulating pedagogy & enabling environment. The latest audio-visual aids and multimedia technology enables the Faculty members to have interactive sessions. Classroom learning is meant primarily for theoretical and conceptual input & consolidated by combining lectures with Case methods and Group Discussion for group learning . Extensive use of laptops is made by students in the well networked class rooms.