Comprehensive opinion analysis on recent QUAD formation in Indo Pacific region using Twitter corpus

Divyansh Singh Panwar
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
divyansh8june@gmail.com

Samarth Garg
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
samarthg1299@gmail.com

Shailender Kumar
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Abstract— **India is known for its highly disciplined foreign policies, strategic location, vibrant and massive Diaspora. India envisages enhancing its scope of cooperation, trade and widens its sphere of relations with the Pacific. As a result, the world is witnessing the rise of Indo-Pacific ties. Before the 1980's the keystone of the universe was called the Atlantic, but now a radical shift to the east is noticed by the term "Indo-Pacific". In this respect, a recent development occurred as a partnership, the Malabar exercise, in the waters of the Pacific and the Indian Oceanic region, which supports and proclaim free, open, and comprehensive Indo-Pacific and stays committed on a rule-based order. Considering the mass inclusion of people on social media platforms and contemplating their opinion has been of much interest in the research. Understanding and categorizing a large dataset of ideas into positive, negative, and neutral aspects is challenging. Our motive is to elucidate how the world, especially the citizens, is welcoming this strategic and influencing partnership. Also, Could social media influence the meaning of partnership? Our paper is divided into two aspects, one dealing with the comparison of various techniques and the other telling opinions of people regarding Indo-pacific relations. We used a geo-specific dataset obtained in various languages. Also, we applied a blend of various ML and DL techniques, feature extraction models, and opinion-based classification, which not only gives an analysis of opinions regarding "Indo-pacific" but also provides an in-depth insight to the comparison of various state-of-the-art sentiment analysis techniques, which is provided as a continuation of earlier review presented on sentiment analysis. The word-cloud visualization system assists people in understanding the changes of public sentiment reactions better.**

Keywords — Data Mining, Sentiment Analysis, Deep Learning, Machine Learning, Artificial Intelligence, Twitter, QUAD, Malabar Naval Exercise, Indo Pacific.

I. INTRODUCTION

Opinion mining refers to the behavioural study of attitudes and opinions of masses regarding an incident or an event that occurred in real-time. Artificial Intelligence, from its beginning, constructed an exceptional commitment in giving practical answers for fundamental cultural human issues under various contours, including NLP, where algorithmic and semantics procedures are utilized to help the computer to perceive and create the ideal outcome. [18] In this present digital era, social media platforms assume a significant part in everybody's daily life. Twitter gives microblogging service a focal share-point where individuals express their beliefs and perspectives on various issues ranging from personal to world politics. According to a report, 456000 tweets were tweeted every minute on average by numerous users worldwide in 2018. [1]

According to various claims, Twitter has a vast network of 330 million monthly users accessing it through the mobile application, website, or SMS.[2] Twitter stays a priority among researchers working in the opinion classification in contrast to its other current conventional partners. It is noticed that Facebook has a more extraordinary user database than Twitter; over there, the data is often diluted with videos, gifs, memes, photos, which doesn't pose usage to researchers. Subsequently, Twitter is seen as the ideal form of thoughts and musings articulation platform. Apart from this, Twitter data is loaded with necessary information only due to its word limit clause. This renovation of data could be an outcome of the recently occurred event. Such a new development occurred recently in world politics, surging the idea of Indo-pacific relations.

India has been a beautiful blend of a different culture, religion and geography. Though India is not aligned to any particular military power, it has maintained affable military & strategic relationships with the majority of friendly countries. Such vibrant relations are also developing with the Pacific under Indo-Pacific relations, which is a recent concept. India maintains warm relations with all the countries but at the same time maintains its Non-alignment attitude. Indo-pacific is a recent term that came into the picture about a decade ago, its rise has been significantly increased in the current time when the whole world started talking about the Indo-Pacific. In that respect, a recent development occurred as an alliance, called QUAD, of four world leaders namely India, Japan, US and Australia to boost the all-time high trade, and increase military and strategic partnership. As a perpetuation in that scenario, Australia agreed to join the Malabar exercise. The Malabar exercise, in the waters of pacific and Indian Ocean, tends to support and proclaim free, open, and comprehensive Indo-Pacific and stays committed on a principles based worldwide request. Hence this development has provided researchers and data scientists a huge chance to explore sentiment analysis. Through this study our motive is to elucidate how world especially the citizens are welcoming this strategic and influencing partnership using various opinion extraction techniques. Thus, this study gives an insight of two aspects, one dealing with extraction of opinions of citizens regarding the new relations and another presenting a thoughtful comparative study of various categories of sentiment analysis techniques namely, lexicon based, machine-learning and deep-learning techniques applied over various feature extraction techniques. This whole study is presented as a continuation of a literature review of numerous sentiment analysis techniques [1], which comprehensively describes currently

state-of-the-art techniques and various metrics used by researchers[1].

This work contains VI sections, section I introduces the paper and gives an insight of the work. Section II deals with the background and procedures opted in the existing studies in order to have a better grasp of the sentiment analysis area and the techniques involved in it. Section III discusses the methodology adopted by us to analyze the opinion in the domain of interest and compares them to various models. It deals with the implementation techniques in the context of the sentiment analysis. It gives a deep insight into our data collection. Section IV shows the obtained results, which will help us visualize the opinions obtained from the dataset. We also analyze the various algorithms used with the help of parameters such as accuracy and time taken. An elaborative and gist comparison is elucidated for better understanding. Section V provides the conclusion by critically analyzing the algorithms used in terms of their significance. We also discussed the future improvements in this field. Last section VI includes the references of the work.

II. BACKGROUND

In recent years, the analysis of tweets has received special attention by researchers, and a number of recent documents published confirms enthusiasm about the sentiment classification of tweets among the scientists.[3] The sentiment classification is often applied on a pre-labeled dataset either available such as used in [4], [5],[6] and from different sources or real-time dataset collected via twitter libraries [7], [2] ,[8], [9], [19] and so on . Thus, a comprehensive and exhaustive literature study is needed to be done in order to know world dynamics and state-of-the-art techniques.

Literature Review: As told, this work is a continuation of a systematic literature review performed on sentiment classification involving social media platforms[1].It highlights various techniques used to analyze the sentiments, along with additional information of metrics used by researchers, dataset availability, short-comings and detailed discussion.

The studies belong to different contours on the basis of approaches they have opted to analyze opinions. They may be categorized into broadly three spaces namely, lexicon based, ML and DL based techniques.

Hitesh et al., [2] aims to analyze the sentiments of Indian citizens regarding General elections, 2019. Authors used about 35,000 live tweets gathered using twitter API. They had used Tf-Idf, word2vec, and BoW as feature extraction techniques over Random Forest classifier to obtain the result. The paper shows importance of word embedding while dealing with NLP. [2]

Similarly, Yu Lim et al., [10] used Logistics Regression ML classifier to analyze the sentiment of a pre-labeled sentiment140 dataset. They used LSA Tf-Idf, ELMo, and word2vec for text representation among which word2vec provided the best results probably due to limited corpus size. [10]. LR is also used in [4].

Chandra et al.,[4]applied many state-of-the-art machine learning techniques such as bernoulli classifier, SDG, Linear SVC, NuSVC and Logistics Regression. Dataset was collected from twitter APIs based on Rafale and Indian military weapons, IMDB movies review, Bitcoin tweets and First GOP debate.[4] Another notable machine learning techniques are SVM shown in [11],[12] and [3], MAXENT [11], Multinomial Naive Bayes [4] , decision tree[13] and KNN[12].

Hwan Khun et al.,[8] analyzed the sentiment of the public when the US banned Huawei amid the US-China trade war. They used lexicon based approach i.e., SentiWordNet to gather opinion using million tweets collected from twitter. They observed a net negative sentiment against Huawei banned around the world. [8]

Ranjan et al.,[9]used the Sentiment dictionary to predict the growth of Indian telecom sector according to subscriber addition. They relied heavily on data from real-time tweets collected using twitter library and subscriber addition data from coai.com.[9]They represented the results in terms of overall sentiment score among which Jio has highest score.[9] Such analysis would help Indian telecom sector to change their policies for better performance.

Similarly, El Rahman et al.,[11]and Harjule et al., [6] used lexicon based models to classify the dataset. [6] used famous pre-labelled Sentiment140 dataset and Crowdflower's data for everyone library.[6]

Advanced Approaches include the involvement of Deep-learning based techniques. Zeeshan Ansari et al.,[14] used RNN based LSTM to extract opinions from 3896 real time tweets generated using twitter API. Their study focuses analysis of Political sentiments orientation of Indian political parties and Indian Leaders.[14] They used Tf-Idf for feature extraction and also used some classical ML classifiers such as SVM, Decision tree, LR, and RF[14].

Khan et al.,[15] presents Hadoop based RNN method over Hadoop framework. Here data was reduced using Map Reduce models. The accuracy obtained was higher as compared to existing methods. Hence we can conclude that hadoop based analysis could be efficient in twitter sentiment analysis [15].

Harjule et al.,[6] also used LSTM technique and compared with both lexicon based SentiWordNet and ML techniques such as LR, SVM, and MNB. They found that RNN have the highest accuracy followed by ML and lexicon respectively.

As noticed almost all the research is focused upon major events occurred in the recent past. The analysis of sentiments is based upon recent development around the world or a massive event which not only moulds world politics but also influences mankind.

Singh et al.,[7] found positive sentiments of citizens regarding Motor Vehicle Act, 2019 even new rules were a burden for people [7]. Similarly, Sahu et al.,[5] analyze the relationship between tweets generated by POTUS and his approval rating using sentiment analytics. Tiwari et al.,[13] analyze the public sentiments after demonetization of 500 and 1000 rupee notes. Other domains of interest where analysis of

sentiment plays significant roles include IPL game[16], Indian General elections of 2019 [2], US-China trade war [8], Rafale induction[4], real-time customer experience[9] and so on.

B. Twitter Data Extraction:

Using the obtained required Access token Secret, API secret, Access token, and API key, user is able to extract tweets using Tweepy python library.
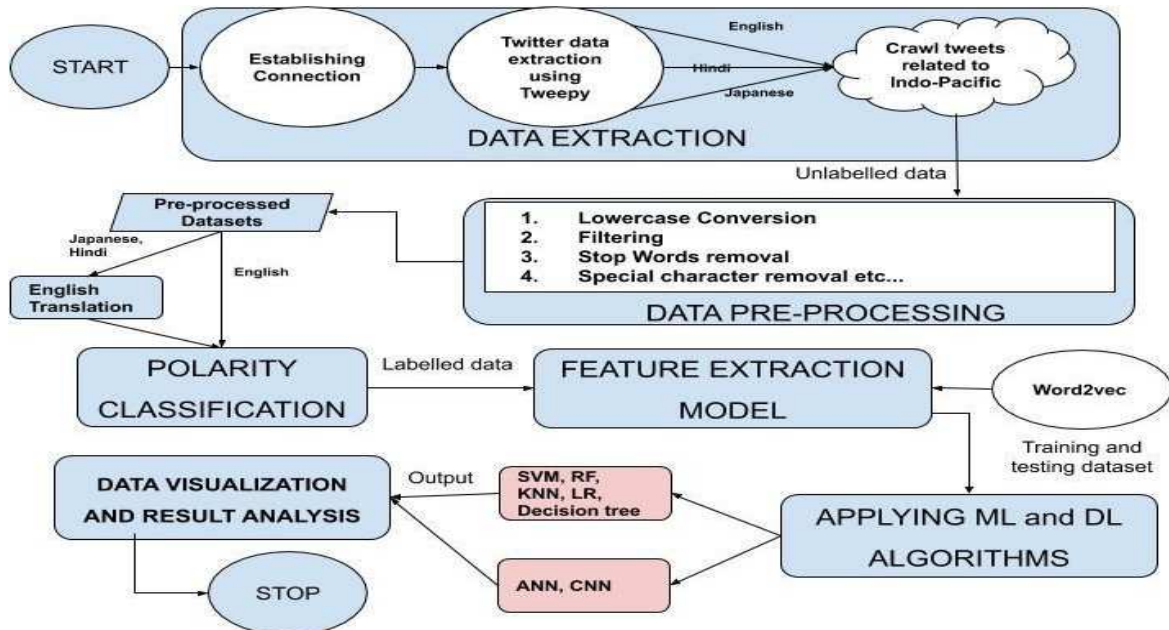

Fig 1. Architectural overview of proposed system

Therefore, information gathered from the analysis of tweets could help the companies, governments and social organizations to better ponder mass public opinion in real-time and make decisions and plans accordingly. The reaction of people could also be noted through the analysis.

III. PROPOSED WORK

It is self-evident that analysis of sentiments in the international relations hemisphere is always a matter of research and played a crucial role in determining world politics. The proposed system mainly consists of 5 steps: Dataset collection, Dataset pre-processing, feature extraction, sentiment classification, and visualization[8]. The architectural overview for the complete process beginning from tweets collection till obtaining the desired results can be observed from Figure 1 below. Stepwise procedure is explained below:

A. Preliminary Link:

1. Acquiring Twitter API keys: Firstly, a user needs to visit developer.twitter.com to create a developer account. Afterward, he needs to login to obtain private keys such as Access token Secret, API secret, Access token, and API key, which are used to get Twitter APIs.

2. Filling the application: The user is required to fill the application details to create an Application.

3. Obtaining private Keys: Then the user is required to go on the "Keys and Access Tokens" tab, to access the information of API secret and API Key. User is required to create the Access token secret and Access token under the Access token section. The user can note it down for further usage.[2]

• Using Twitter library: Tweepy, which is a python based library, helps to extract the tweets directly from Twitter through the twitter API. While establishing connection, a PIN is generated. Thereby, using that PIN one can access Twitter API and extract tweets.[2]

• The tweets were extracted using the a function "searchTwitter" and based on two search words; those are #QUAD, and #MALABAR.

C. Dataset Overview:

The tweets collection was started by passing coordinates of the desired location as the focus of the study is over the QUAD countries only. As the Malabar exercise was scheduled in two phases in November, thus data collection continued for over two months i.e. from 17/Nov/2020 to 13/Jan/2020. The dataset collected was categorized and stored in a geo-localized manner. The categorization of the dataset was on the basis of language i.e.. It was divided into three parts, namely English, Japanese, and Hindi. Over the month, 35,000 annotated tweets were collected, after which the following steps of pre-processing, analysis, and visualization were applied. used. The link on which dataset is available on: https://ieee-dataport.org/documents/quad-dataset [17].

The data preparation is the keystone and most cumbersome stage of the implementation of the study. As the dataset is considered as the gist of the project, the higher accuracy of this data will lead to a more reliable output.

The dataset includes following attributes; username, location, tweet, date, tweet coordinates. Figure 2 provides a snapshot of all three dataset vis-à-vis Japanese, Hindi and English.

user	index_col	location	tweet	Date	Tweet Coordinates
TV9Bharatvarsh	1	Noida, India	अरब सगर म मलबर यदधभयस क दसर फज आज स शर quad दश...	2020-11-17 03:06:00	NaN
TV9Bharatvarsh	2	Noida, India	दखए दशदनय क हर बड खबर लकल to गलबल top 100 Hloc...	2020-11-17 03:19:32	NaN
TV9Bharatvarsh	3	Noida, India	समदर म चन क खलफ quad दश क तयरthesamirabbas gen...	2020-11-17 07:00:39	NaN
SDeviputra	4	Jodhpur, India	tv9bharatvarsh अरब सगर म मलबर यदधभयस क दसर फज ...	2020-11-17 10:06:59	NaN
TV9Bharatvarsh	5	Noida, India	मलबर एकसरसइज स चन क हलत खरबदखए फल फइनलmalabarw...	2020-11-17 11:24:37	NaN
user					
jagadishrjain	1	Hyderabad	rrrameshrrr 50 of the quad japan australia and...	2020-11-15 09:00:05	NaN
hellfire_81	2	NaN	coolquad riplegic military grouping httpstcocx...	2020-11-15 09:09:46	NaN
Savitri4Ever	3	NaN	hellfire_81 coolquad riplegic military groupin...	2020-11-15 09:19:10	NaN
user					
Ln__o0	1	ゲームしてる人だけ400人	yuk1dqma_quad 私の2リグか4リグ更新してもろて	2020-11-30 15:03:14	NaN
QUad_R_ANgle	2	////りがくぶ・くりかえす・ふじまめ ←一言メッセージ	samemizu ゲーム開発初めて6か月ついにアプリをリリース出来ましたくっつきパズル のび...	2020-11-30 15:03:40	NaN
QUad_R_ANgle	3	////りがくぶ・くりかえす・ふじまめ ←一言メッセージ	今日からペンパアドベントか	2020-11-30 15:04:40	NaN

Fig 2. Snapshot of multilingual dataset

D. Dataset Pre-processing:

For a systematic analysis data preprocessing is a crucial stage. It changes the crude information into refined data. The data obtained in real-time frame has such an uproar and is frequently deficient, conflicting, and ailing in certain practices or patterns. In this stage, a python based NLTK library was incorporated. The data was passed through following steps:

- Convert the data into lowercase and remove the twitter handles or other usernames.

- Remove the links, stop-words and extraordinary symbols since they don't contribute to data interpretation.

- Word-tokenization or Lemmatization: this turns the data into defined entities.

- Spelling Correction: It will correct the wrongly spelled words.

These steps will improve the system exactness and decrease the required computational time.

E. Feature Extraction

The obtained unlabelled dataset was labeled using the textblob sentiment classifier. The motive behind the labeling is to make our dataset suitable for supervised ML classification. Textblob is a python library that provides simple APIs to classify the sentiments. We imported NaiveBayes analyzer for classification. For a better sentiment analysis model, we have to transform our dataset into a more straightforward format. We have to extract all our data features and represent them into the numeric format, usually called vectors. Word2vec, a DL neural network-based feature engineering technique, is used to extract information from a raw corpus and train different classification models. Word2vec uses two basic methods, CBOW and Skip-gram technique, to map a target variable with the available words. It outweighs other classical TF-IDF and BoW in terms of the significant reduction in dimensionality, incorporation of contextual analysis.[2]

F. Sentiment Classification

After polarity classification and applying respective feature extraction models, the dataset is split into training and testing datasets in 80 to 20 ratio, respectively. The purpose behind is training our model using a training dataset and further validating the trained model using a testing dataset. For the classification of sentiments and analysis of results, different ML classifiers were used. In addition to ensuring accuracy and reliability of outcome, state-of-the-art deep learning techniques were also incorporated for analysis of opinion. SVM and LR algorithms, which are ML-based linear supervised classifiers, were used for sentiment classification. They were compared with another set of ML-based non-linearsupervised classifiers, including Random Forest, KNN, and Decision tree. Apart from ML algorithms, DL based ANN andCNN were used for classification, and obtained results were compared.

G. Data Visualtization

The results are used to validate and visualize the opted methodology. Precision, F1-score, Accuracy and recall assist in depicting the results. They are termed as key performance indicators[1]. Other creative methods such as WordCloud, geographic visualization also help to envision outcomes. A word cloud shows the significance of each word in the dataset. The word's size and shade signify how much a specific term is being utilized in the dataset, as appeared in Figure 3.

Fig 3. Wordcloud visualization of corpus

IV. DISCUSSION AND RESULT ANALYSIS

The 'Indo-pacific' dataset consists of 29,499 tweets upon which various machine and deep learning techniques were applied. The model defines (+1) for the positive tweets and (-1) for the negative tweets. Those tweets which stood neutral had (0) as their value assigned. The system shows the agreement's sentiment analysis result using a bar graph, as shown in figure 4. The overall positive tweets were 12183, negative was 4131, and neutral were 13185.
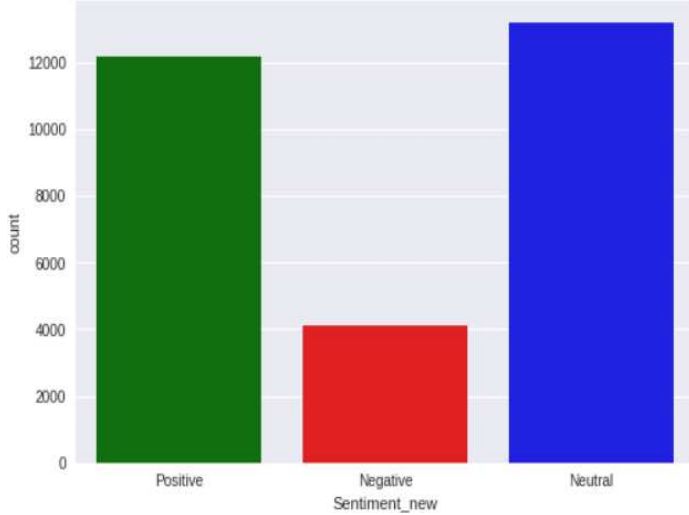

Fig 4. Sentiment classification

Considering the individual performance of each model, SVM outperforms all other machine learning models and deep learning models with overall accuracy of 98% followed by random forest, KNN, decision tree, and logistics regression respectively. The results could be referred from the table 1.

| S NO.	Approach	Accuracy	Time taken (sec)
Machine learning			
1	SVM	98%	20.83
2	Random Forest	96%	24.85
3	KNN	94%	38.75
4	Decision Tree	93%	11.99
5	Logistic Regression	89%	2.31
Deep learning			
1	CNN	94.20%	298
2	ANN	91.06%	202

Table 1. Comparison of various approaches

Logistics regression takes least time to train. It took 2.31 seconds the model to train and show the results. 94.20 and 91.06% accuracy is observed for CNN and ANN respectively, which is higher than its logistics regression. Figure 5 represents the confusion matrix of the test dataset used by the ANN model.
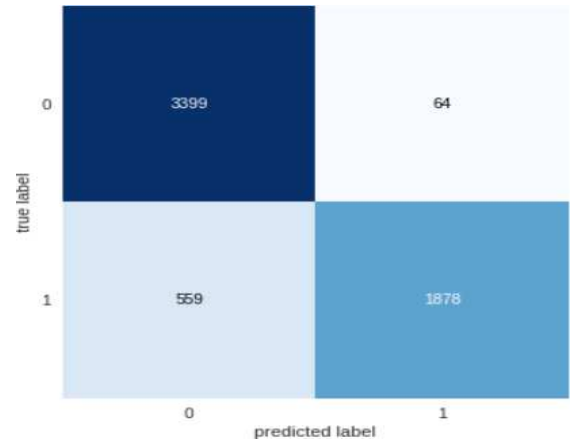

Fig 5. Confusion matrix for ANN model

The scatter plot is a dot based representation, which is used to observe the relation between the variables based on two axis. Figure 6 gives a glimpse of relationship between positive, negative and neutral sentiments. It is clearly visible that negative tweets are distinguished from the positive tweets from the neutral tweets in between.
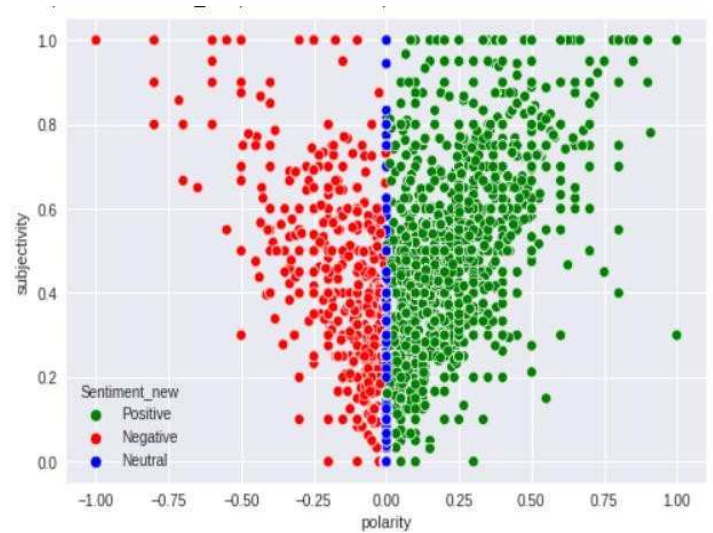
Fig 6. Labeled corpus Scatter Plot representation

Figure 7 tells the accuracy of CNN model after each epoch of training.
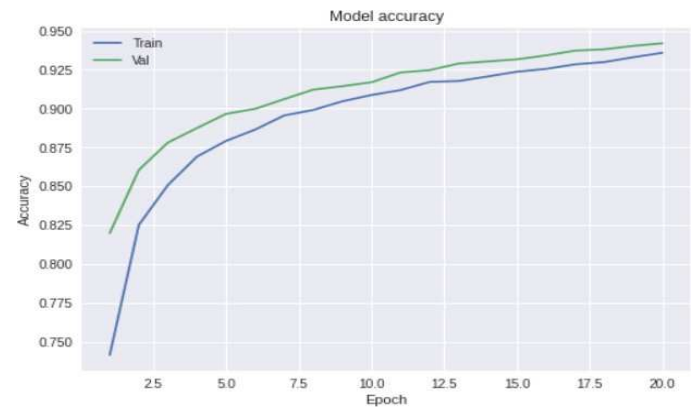
Fig 7.CNN model accuracy

V. Conclusion

The analysis of social media platforms provides reliable and rich information. In this paper, an exhaustive comparison of numerous ML and DL techniques is presented. It was inferred that SVM, which is a linear ML classifier, has outperformed all other ML and DL algorithms. It is observed to have 98% accuracy, which is admirable for a machine learning algorithm. CNN is observed to have 94.20% accuracy, which is also an excellent accuracy keeping the limitation of the dataset's size in view. It is apparent that as the size of the dataset increases, the accuracy of DL algorithms also increases [4]. Thus, considering the size of the dataset, CNN accuracy is quite exceptional.

Along with the comparison, a comprehensive analysis of residents' sentiments from QUAD nations is performed over the recently developed "Indo-pacific" relations. From the results acquired, It is evident that the people have accepted the idea of free Indo-pacific. This is marked as an onset of changing economic and security dynamics of the world.

For future purposes, more data could be collected from the recent developments hence providing more detailed results. New domains could also be explored, such as hatred speech detection, i.e., mitigation by analyzing the tweets of hate spreaders, sentiment mining of various patient tweets to ascertain their mental disturbance level. We have focused our study on QUAD only, but the study could be extended to India's foreign relations and policy. Most importantly, the work is carried out in ML and DL based techniques. But the work which we incorporated is computationally expensive and slow in terms of runtime. Thus, new approaches could be identified which improve the issues mentioned earlier and various metrics.

VI. References

[1] S. Garg, D. S. Panwar, A. Gupta, and R. Katarya, "A Literature Review On Sentiment Analysis Techniques Involving Social Media Platforms," *2020 Sixth Int. Conf. Parallel, Distrib. Grid Comput.*, pp. 254–259, Nov. 2020.

[2] M. S. R. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam, and S. Kumari, "Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model," in *ICCT 2019*, IEEE, 2019.

[3] N. Pawar and S. Bhingarkar, "Machine Learning based Sarcasm Detection on Twitter Data," in *5th Int. Conf. on Communication and Electronics Systems (ICCES)*, IEEE, 2020.

[4] Y. Chandra and A. Jana, "Sentiment analysis using machine learning and deep learning," in *Proceedings of the 7th Int. Conf. on Computing for Sustainable Global Development*, IEEE, 2020.

[5] K. Sahu, Y. Bai, and Y. Choi, "Supervised Sentiment Analysis of Twitter Handle of President Trump with Data Visualization Technique," in *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, IEEE, 2020.

[6] P. Harjule, A. Gurjar, H. Seth, and P. Thakur, "Text Classification on Twitter Data," in *Proceedings of 3rd Int. Conf. on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, ICETCE 2020*, IEEE, 2020.

[7] N. A. Singh, "Sentiment analysis on motor vehicles amendment act, 2019 an initiative by government of India to follow traffic rule," in *Int. Conf. on Computer Comm. and Informatics,* IEEE, 2020.

[8] N. H. Khun and H. A. Thant, "Visualization of Twitter Sentiment during the Period of US Banned Huawei," in *ICAIT*, IEEE, 2019.

[9] S. Ranjan, S. Sood, and V. Verma, "Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies," in *Proceedings - 4th Int. Conf. on Computing Sciences, ICCS 2018*, IEEE, 2019.

[10] Y. Q. Lim, C. M. Lim, K. H. Gan, and N. H. Samsudin, "Text Sentiment Analysis on Twitter to Identify Positive or Negative Context in Addressing Inept Regulations on Social Media Platform," in *ISCAIE 2020 - IEEE 10th Symposium on Computer Applications and Industrial Electronics*, IEEE, 2020.

[11] S. A. El Rahman, F. A. Alotaibi, and W. A. Alshehri, "Sentiment Analysis of Twitter Data," in *2019 International Conference on Computer and Information Sciences, ICCIS 2019*, IEEE, 2019.

[12] S. M. Parikh and M. K. Shah, "Classification of sentiment analysis using machine learning," in *Lecture Notes on Data Engineering and Communications Technologies*, 2020.

[13] S. Tiwari, A. Verma, P. Garg, and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," in *6th Int. Conf. on Adv Computing and Comm. Systems, ICACCS 2020*, IEEE, 2020.

[14] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," in *Procedia Computer Science*, ScienceDirect, 2020.

[15] M. Khan and A. Malviya, "Big data approach for sentiment analysis of twitter data using Hadoop framework and deep learning," in *International Conf. on Emerging Trends in Information Technology and Engineering, ic-ETITE 2020*, IEEE, 2020.

[16] V. Prakruthi, D. Sindhu, and S. Anupama Kumar, "Real Time Sentiment Analysis of Twitter Posts," in *2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018*, IEEE, 2018.

[17] Divyansh Singh Panwar, Samarth Garg, "QUAD DATASET", IEEE Dataport, doi: https://dx.doi.org/10.21227/jy92-e424.

[18] H. Rehioui and A. Idrissi, "New Clustering Algorithms for Twitter Sentiment Analysis," in IEEE Systems Journal, vol. 14, no. 1, pp. 530-537, March 2020.

[19] H. Liang, U. Ganeshbabu and T. Thorne, "A Dynamic Bayesian Network Approach for Analysing Topic-Sentiment Evolution," in IEEE Access, vol. 8, 2020.