International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis of Yelp Reviews using Various State of the Art ML Models

Ambica M. Kalcoor

Visvesvaraya Technological University

Abstract : Today, prospects consider online reviews as an important source for product information before proceeding with the purchase. This increasingly connected world has allowed a single voice to be heard by millions, which makes a customer's online review critical for product sales. This research lays emphasis on Sentiment Analysis to extract user opinion from the crowd-sourced review forum, Yelp. An attempt is made to generate a user specific sentiment analysis aside from a general analysis in this paper. Further, user-specific extraction methods are implemented to analyze linguistic features associated with features such as user's writing style and vocabulary. These features are utilized to train models and make sense of the noisy dataset acquired for the research. Models such as Multinomial Naive Bayes, Gradient Boosting, K-nearest Neighbors, and Adaboost were compared and contrasted to determine the one that can offer the highest classification accuracy rate. In the end it was found that Adaboost achieved the 87.46% classification accuracy surpassing all the other models.

1. INTRODUCTION

Our opinions are inherently shaped by the perceptions of others and how they evaluate the world. This underlines the reason why 'word of mouth' has been an important factor for sales metrics. In the past, the effect of word of mouth has always been limited to friends and relatives. Now, user-generated feedback on online forums has become an important channel of product information making the opinion of one customer accessible to thousands of potential buyers [1]. Therefore, it is rational for users to rely upon testimonials and reviews before making a purchase online. As a matter of fact, they are accustomed to use online reviews information as a basis to judge whether they purchase. [2] The seamy side to this process involves wading through the vast data exploding all over the web and discriminate between relevant and irrelevant information.

An interesting work surrounding this topic includes "Manipulation of online reviews: An analysis of ratings, readability, and sentiments". It delves into understanding to what extent these online reviews are truthful 'user-generated' reviews or merely reviews provided by vendors interested to push the sales of products [3]. It is also observed that readers are often interested in a detailed opinion analysis instead of general sentiment towards a product[4].

As a result, studying the characteristics of user-generated reviews has now emerged as a potential research area specifically for opinion mining. Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [5].

This research paper leverages sentiment analysis to classify Yelp Review sentiments as positive or negative. Further, user-specific extraction methods are implemented to analyze linguistic features associated with features such as user's writing style and vocabulary. These features are utilized to train models and make sense of the noisy dataset acquired for the research. The user information and general information are passed on to two different models. This process helps in generating two individual text representations with user attention or generalized analysis. Various models were used to perform sentiment analysis and, in the end, the best model was ascertained from the lot.

2. DATA DESCRIPTION

Our main focus for sentiment analysis will be on Yelp's review dataset. Yelp is the popular crowdsourced review forum that embodies millions of reviews covering a wide spectrum of businesses including restaurants, salons, and even plumbers. It is because of this wide range of audience that Yelp is an ideal playground for sentiment analysts.

The reviews are aggregated into files in one JSON-object per-line format. The dataset contains 6 Million reviews given by 1.5 Million users for around 188595 businesses. For the purposes of this project, we have considered reviews with stars 1,2,4 and 5. Also, we only consider those users who have given 500 or more reviews for user-specific sentiment analysis to generate better accuracy across the user-specific models.

In Fig. 1, the first few rows of the dataset is shown. It is represented in .csv format after processing the JSON files and extracting the necessary data.

business_id	date	review_id	stars	text	user_id	cool	funny	useful
iCQpiavjjPzJ5_3g	2/25/2011	x7mDIiDB3jEiPGPH	2	The pizza was okay. Not the b	msQe1u7Z_XuqjGoqhE	0	0	0
pomGBqfbxcqPv	11/13/2012	dDl8zu1vWPdKGih	5	I love this place! My fiance An	msQe1u7Z_XuqjGoqhE	0	0	0
jtQARsP6P-LbkyjI	10/23/2014	LZp4UX5zK3e-c5ZG	1	Terrible. Dry corn bread. Rib t	msQe1u7Z_XuqjGoqhE	1	1	3
elqbBhBfElMNSr	2/25/2011	Er4NBWCmCD4nM	2	Back in 2005-2007 this place w	msQe1u7Z_XuqjGoqhE	0	0	2
Ums3gaP2qM3W	9/5/2014	jsDu6QEJHbwP2Blc	5	Delicious healthy food. The st	msQe1u7Z_XuqjGoqhE	0	0	0
vgfcTvK81oD4r5(	2/25/2011	pfavA0hr3nyqQ61	1	This place sucks. The custome	msQe1u7Z_XuqjGoqhE	0	0	2
AxeQEz3-s9_1Ty	10/10/2011	brokEno2n7s4vrw	5	If you like Thai food, you have	msQe1u7Z_XuqjGoqhE	0	0	1
zdE82PiD6wquvj	4/18/2012	kUZWBVZvhWuC8	5	AMAZING!!! I was referred he	msQe1u7Z_XuqjGoqhE	0	1	0
EAwh1OmG6t6p	2/25/2011	wcqt0IlI88LEcm19l	4	Ribs = amazing2 hour wait tim	msQe1u7Z_XuqjGoqhE	0	0	0
atVh8viqTj-sqDI:	11/9/2012	LWUtqzNthMM3vy	2	Food is pretty good, not gonn	msQe1u7Z_XuqjGoqhE	1	2	1
yFumR3CWzpfvT	6/15/2016	STiFMwwZz31siPY	5	I have been an Emerald Club r	TlvV-xDhmh7LCwJYXkV	0	0	0

Fig. 1. Example of rows in dataset

A. Data Exploration

In this section, the data exploration in the project is presented with figures giving a rich insight of the data. These would also help in deciding the data pre-processing and feature extraction methods. The distribution of ratings can be observed in the figure below. It can easily be seen that the data is highly unbalanced.

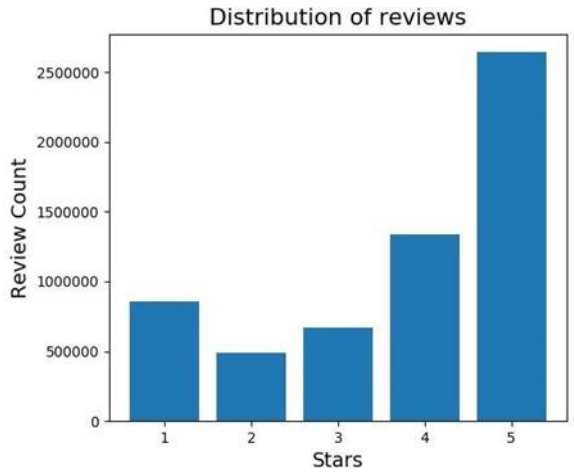

Fig. 2. Rating Statistics (in stars)

In Figure 3, the distribution of average of stars per user is shown. We can see general spikes in both directions, with data neutralizing for more number of users.

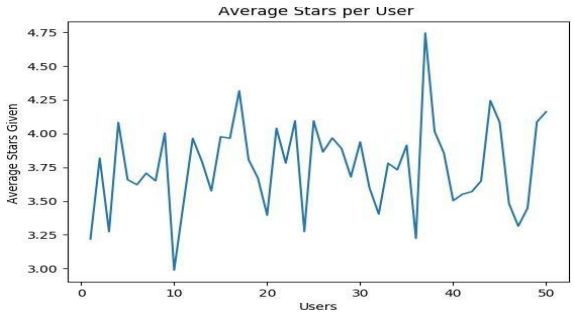

Fig. 3. Average stars per user

In the Figure 4, the *histogram* of text length distributions for each star rating is shown. The distribution helps us decide which metrics would be best suited for sentiment analysis.

In the histogram, the 4 and 5 starred reviews generally have the highest word count. Positive reinforcements would thus be easy to learn from this dataset. It also points towards the probability that the dataset is more positively inclined as compared to negative examples that occur in the training dataset.

We decided to omit all 3-star reviews to more accurately balance our dataset and not include any ambiguous reviews that do not add value to the models.

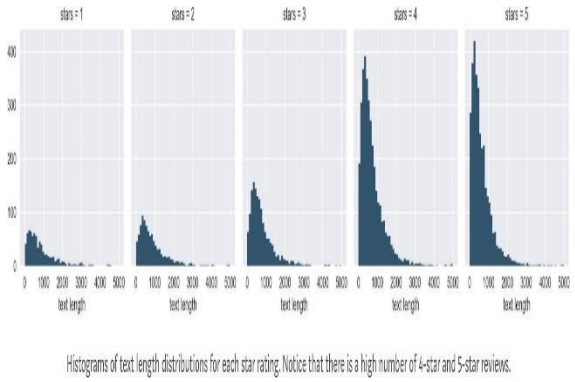

Fig. 4. Histograms of text length distributions for each star rating

In order to develop a user specific model, we decided to go with users who have reviewed establishments on Yelp more than 500 times. The reason for deciding this number stems from the data exploration and also from the belief that such users will tend to give more genuine reviews as compared to others. It will also help us test our model based on the particular style of that user so that we don't have to generalize the weights for a review word across all users.

3. METHODOLOGY

To perform both classic and user-specific sentiment analysis, a series of steps need to be followed. At first, various pre-processing methods were applied on the reviews. Features were extracted from these reviews and fed into the different models. These steps are crucial as they form the foundation of modelling.

1) Data Pre-processing: In this step, data is converted to .csv format, and standard approaches like stop-word removal, data cleaning etc. are employed along with approaches specific to our dataset, like selection of users and reviews to model user-specific sentiment analysis, removal of neutral reviews to create a more robust dataset and so on.

2) Feature Extraction: Here, we perform TF-IDF vectorization to extract the features. The classic method involves applying TF-IDF on the corpus of all reviews. Along with this, we also apply TF-IDF vectorization with respect to users. This user- specific approach gives us a better understanding of the importance of the words in a given user's language.

3) Classification: Finally, we choose different classifiers and train a model for each user. The accuracy is computed as a combination of accuracy measures across all the models. This accuracy is compared with the classification accuracy obtained from classic method which involves training one model for a given classifier.

These processes are explained in further detail in the following sections. The methodology used by us is also depicted in Figure 5

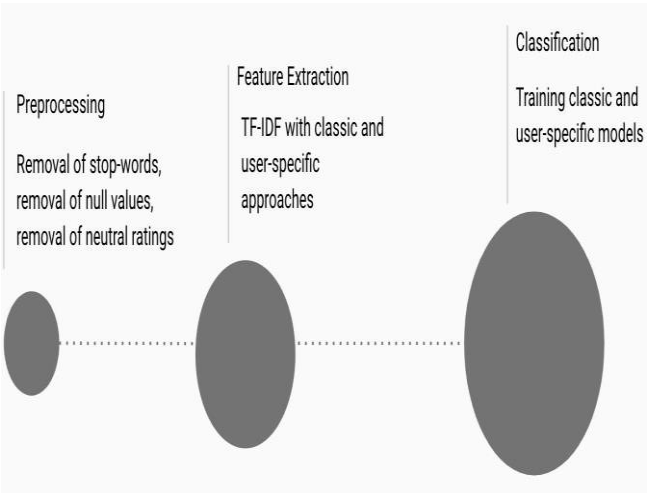

Fig. 5. Methodology

4. DATA PRE-PROCESSING

In this section, different data pre-processing techniques and data cleaning techniques are discussed. Few of these steps are the standard pre-processing procedures that need to be applied for any sentiment analysis and model building task. Apart from these, we also apply a few methods specific to our dataset. This helps us in filtering out irrelevant information and preserving the most useful representation of the reviews.

A. Extraction of data from JSON format

Yelp's dataset is provided in a JSON format that is represented in Figure. 6.

The following data attributes were gleaned from the dataset:

1) business id: ID of the business being reviewed
2) date: Day the review was posted
3) review_id: ID for the posted review
4) stars: 1 to 5 rating for the business
5) text: Review text
6) user_id: User's ID
7) cool/useful/funny: Comments on the review, given by other users

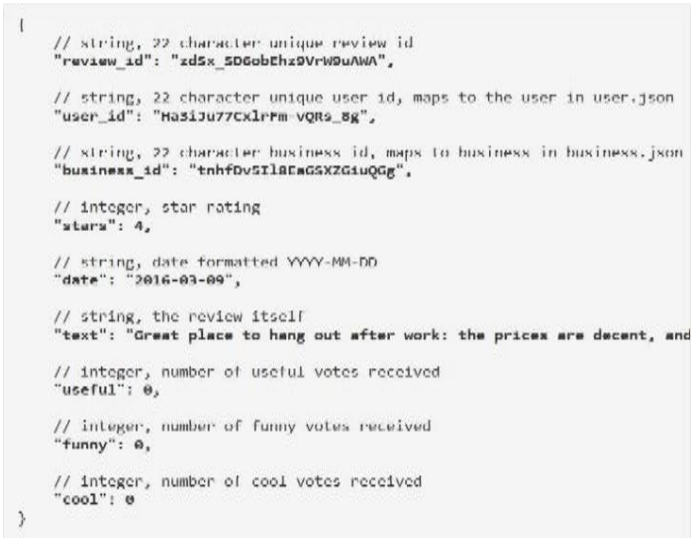

Fig. 6. JSON representation of data

TABLE I

EXAMPLES OF WRONG SPELLINGS

Wrong Spelling	Correct Spelling
Refrigerator	Refrigerator
Rechargeable	Rechargeable
Shoppe	Shop
Refrigerator	Refrigerator

The JSON file is converted to CSV format to improve readability and make it compatible with programming tools.

B. Eliminating Stop words

These words do not confer meaning by themselves, but are frequently occurring in a sentence. These words are extremely common and would have a very little value of their inclusion in the vocabulary. The figure below depicts a representation of the same.

a an and are as at be by for from
has he in is it its of on that the
to was were will with

Fig. 7. Examples of stop words

C. Spelling correction

As the nature of online reviews is usually informal, the data is expected to contain spelling errors. Table 1 is composed of misspelled terms in the dataset. It is crucial to correct the spellings as a consistent dataset for future analysis is necessary.

D. Removal of neutral reviews

3-star reviews present a neutral sentiment and are susceptible to misclassification depending on system bias. We thus decided to remove these reviews from the training set for the purposes of our experiment.

E. Identification of users

For user specific sentiment analysis, we needed a training set for each user with a considerable number of reviews, for the model to be able to identify the style in which the user writes his/her reviews. After data exploration, we decided to choose a subset of 50 users who have written greater than or equal to 500 reviews each.

F. Preserving Useful Reviews

The dataset contains 'useful' as one of the attributes. This is a rating given for a particular review by other users. Using this label, we can filter out the reviews that do not provide us with significant or new information. Also, using the 'date' attribute, we filter out outdated reviews.

5. FEATURE EXTRACTION

The reviews cannot be directly used as features, as the dataset in this project does not translate to numerical features, but to textual data. Text data has to be converted to numerical vectors to fit the machine learning models.

A. Methods to convert Text to Numerical Vectors

1) TF-IDF vectorization: The reviews in the Yelp dataset cannot be directly used as features for the learning models as they represent textual data. Each review needs to be converted to a numerical vector. The models take in these numerical vectors as representation of reviews and learn from them. We have used TF-IDF for extracting features from reviews.

Term Frequency-Inverse Document Frequency (*TF-IDF)* vectorizes given text on grounds of the word's importance in the existing document. Term Frequency (TF) and Inverse Document Frequency (IDF) are computed for every term in the vocabulary. They are and which would be defined below,

Let,

x = Number of times term t occurs in document i

y = Total Number of terms in the document i

$TF = {}^{x}/y$... (1)

Let,

a = Total Number of Documents

b= Total Number of Documents containing term t

$IDF = \ln^{a}/_{b}$... (2)

The TF-IDF feature for document *i* and term *t* would be $TF * IDF$. Hence a particular term can be represented as a vector of the size equivalent to the number of documents, with the respective TF-IDF weights associated with it.

2) User Specific TF-IDF vectorization: When we use the classic method of applying TF-IDF over a corpus containing all the reviews, we do not take the difference in the language of different users into consideration. For instance, User1 uses the word 'good' in his/her reviews a lot more times than User2 does. User1 uses this positive word with much more liberty than User2. So the weights given to this particular word differs between the two users. This difference in treat- ment of words in the language is not reflected in the classic approach. Therefore, we experiment with a user-specific TF- IDF vectorization where TF-IDF is calculated independently for each user. The frequency is evaluated over the corpus of the reviews belonging to a particular user. In this approach, the difference in treatment of words is represented as TF-IDF for a given user is built only using his/her reviews.

6. MODELS

A. Multinomial Naive Bayes

Naive Bayes is a learning algorithm that is frequently employed to tackle text classification problems. The multinomial event model—frequently referred to as Multinomial Naive Bayes or MNB for short—generally has been found to compare favorably with more specialized event models[6].

A multinomial event model is composed of the frequencies of events produced by a multinomial p_i indicating the probability of i's occurrence.

This makes a feature vector x turn into a histogram where x_i indicates the count of event i. The probability of observing the histogram can be described as follows:

$$p(x|C_K) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i} \quad \ldots (3)$$

B. Gradient Boosting

Gradient boosting is a state-of-the-art prediction technique that sequentially produces a model in the form of linear combinations of simple predictors—typically decision trees—by solving an infinite-dimensional convex optimization problem[7]. Given a least-square setting, it can be described with a model F to predict values that can be represented

As $\hat{y} = F(x)$, $\sum_i(\hat{y}_i - y_i)^2$. Here, i resembles a training set of size n that are composed of the the values present in y.

This method presumes an actual value of y and estimates $\hat{F}(x)$ by weighing the sum of functions $h_i(x)$ from weak learners

$$\hat{F}(x) = \sum_{i=1}^{M} \gamma_i h_i(x) + \text{const.}$$

We applied Gradient Boosting on the TF-IDF vectors generated for both the classic and user specific sentiment analysis models. Being an ensemble technique, Gradient Boosting performed much better when compared for both accuracy and recall with Multinomial Naive Bayes.

C. K-nearest Neighbor (KNN)

It is a supervised learning technique that is based on discovering analogous objects from sample groups using a distance metric and evaluating the new unseen data into the same class as that of the majority of neighbors. The nearest neighbor (NN) rule identifies the category of unknown data point on the basis of its nearest neighbor whose class is already known [8]

The algorithm considers that documents can be categorized in the Euclidean space as points. The distance between two points can be computed as follows

$$d(q,p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

The K-nearest neighbours to a review can be computed using distance metrics other than Euclidean distance as well. Some of the popular distance metrics are Mahalanobis distance, Manhattan distance, Minkowski distance, etc. For the given dataset, we found that Euclidean distance was the best measure.

D. AdaBoost

AdaBoost that is an abbreviation of Adaptive Boosting is a boosting algorithm and one of the most important multiple classification methods because of its strong theoretical foundation benefits, highly accurate computations, and simplicity[9]. As an important realization of boosting theory, AdaBoost is extremely easy to implement and keeps competitive in terms of both practical performance and computational cost[10]. It uses the technique to train a boosted classifier that can be defined below:

$$F_T(x) = \sum_{t=1}^{T} f_t(x)$$

Here, f_t represents a weak learner that returns a value predicting object class after being trained by input x. The

weak learners produce an output hypothesis, $h(x_i)$ at every iteration t where a weak learner is attached to a coefficient α_t such that the sum training error E_t fetched from t-stage boost classifier reduces to minimum value.

$$E_t = \sum E[F_{t-1}(x_i) + \alpha_t h(x_i)i$$

$F_{t-1}(x)$ resembles the boosted classifier in the above equation and the weak learner's weight is represented as $\alpha_t h(x)$. In our experiments, AdaBoost was the best performing model for user specific review analysis out of all the models that we tried. Its accuracy was comparable to that obtained using the classic review analysis model generated using Ada Boost.

7. RESULTS

The below are the performance measures obtained for different models. The metrics that have been considered to evaluate the different models used are: Accuracy, Precision, Recall and F1- Score.

TABLE II

PERFORMANCE MEASURES FOR MULTINOMIAL NAIVE BAYES

	Classic	User Specific
Accuracy	88.7%	73.79%
Precision	0.90	0.74
Recall	0.98	1
F1 Score	0.94	0.85

Table II compares the performance of Multinomial Naive Bayes on the classic and user specific models. The accuracy of the user specific model is poor. However, the recall value of the classification model on user specific data becomes 1.

TABLE III

PERFORMANCE MEASURES FOR GRADIENT BOOSTING

	Classic	User Specific
Accuracy	89.5%	85.47%
Precision	0.93	0.87
Recall	0.85	0.94
F1 Score	0.89	0.91

Table III reports the performance of Gradient Boosting on sentiment analysis of the reviews. The table shows that the accuracy of Gradient Boosting on the classic and user specific models are comparable. The F1 score of the user specific model is better than that of the classic model.

TABLE IV

PERFORMANCE MEASURES FOR KNN

	Classic	User Specific
Accuracy	90.1%	84.9%
Precision	0.88	0.86
Recall	0.97	0.95
F1 Score	0.92	0.9

Table IV reports the performance of KNN for classic and user specific models. The precision, recall and F1 scores of KNN is comparable between the classic and user specific models. The best model was obtained for Euclidean distance metric and 60 nearest neighbours.

TABLE V

PERFORMANCE MEASURES FOR ADABOOST

	Classic	User Specific
Accuracy	92.4%	87.46%
Precision	0.96	0.89
Recall	0.98	0.95
F1 Score	0.97	0.92

It can be seen from the above table that the gradient boosting models outperform other preliminary techniques. AdaBoost with Gradient Boosting Decision Tree has given the best results out of all the models experimented.

8. CHALLENGES

Following are some of the challenges we faced during this project:

- Storage requirements for user-specific analysis - As we are building a model for each user, these models need to be stored for predictions in the future. More the number of users involved, more is the storage requirement for all the trained models.
- Minimum requirements for reviews per user - We retain reviews from only those users who have posted greater than or equal to 500 reviews. Such a constraint poses a challenge which is hard to deal with when the data is not sufficient.
- Identifying sarcasm, double negation and user extended vocabulary.

The main challenge of sentiment analysis remains tackling the nuance of a language. For example, the particular challenge of classifying these reviews:

- I do not dislike the food here. *(Negation handling)*
- Sometimes I really hate the ribs at this joint, but I keep coming back for more. *(Adverbial modifies the sentiment)*
- Yes! Please keep us waiting for a salad for 30 mins. You know how much we love to sit and count tiles in this fine establishment. *(Possibly sarcastic)*

9. FUTURE SCOPE

We have identified the following modifications to the existing model to give richer results:

1) Combination of two TF-IDFs - The classic approach and user-specific approach can be combined to generate a much more powerful representation of the data.
2) Combining models of different users - We can combine models for users with similar language instead of building a model per user. This will drastically bring down the storage requirements.
3) We can attempt to predict how genuine a particular user's comment will be based on sentiment classification of past reviews by the same user.

10. CONCLUSION

In this project, a very relevant challenge with regards to user specific and general sentiment analysis was tackled. Many features and meta-data were considered and various ways of achieving user centric sentiment analysis were touched upon. Numerous models were also experimented upon to provide a robust and highly accurate solution. The discussed approaches would perform very well on other sentiment analysis tasks as well.

REFERENCES

[1] Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. Journal of management information systems, 23(2), 149-171.

[2] F., & Fan, P. (2015). Effect of online reviews on consumer purchase behavior. Journal of Service Science and Management, 8(03), 419.

[3] Titov, I., & McDonald, R. (2008, April). Modeling online reviews with multi-grain topic models. In Proceedings of the 17th international conference on World Wide Web (pp. 111-120).

[4] S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision support systems, 52(3), 674-684.

[5] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167

[6] Kibriya, A. M., Frank, E., Pfahringer, B., & Holmes, G. (2004, December). Multinomial naive bayes for text categorization revisited. In Australasian Joint Conference on Artificial Intelligence (pp. 488-499). Springer, Berlin, Heidelberg.

[7] Biau, G., Cadre, B., & Rouvière, L. (2018). Accelerated gradient boosting. arXiv preprint arXiv:1803.02042.

[8] Dhanabal, Subramaniam & SA, Chandramathi. (2011). A Review of various k-Nearest Neighbor Query Processing Techniques. Int. J. Comput. Appl.. 3.

[9] Mazini, M., Shirazi, B., & Mahdavi, I. (2019). Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. Journal of King Saud University-Computer and Information Sciences, 31(4), 541-553.

[10] Sun, K., Lin, Z., & Zhu, Z. (2019). Adagcn: Adaboosting graph convolutional networks into deep models. arXiv preprint arXiv:1908.05081.