Classification Using SVM, Naïve Bayes, KNN, Decision Tree, and Supervised Machine Learning Techniques for Prediction of Wind and Solar Power Outputs in Renewable Energy Systems.

Kondapalli Srinivasa Varaprasad[1], Raju Basak[2], Sourish Sanyal[3], Abhro Mukherjee[4], Arunava Kabiraj Thakur[5]

Techno India University, WB, India[1,2,4] , Alipurduar Govt. Engineering & Management College[3] , Techno Main Salt lake[5]

Abstract

In the era of increasing reliance on renewable energy, accurate prediction of wind and solar power outputs has become crucial for efficient energy management and grid stability. This study explores the application of various supervised machine learning algorithms—including Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbours (KNN), and Decision Tree classifiers—for the classification and prediction of wind and solar power generation. The models are trained on historical meteorological and energy output data to identify key patterns and dependencies influencing power output. A comparative analysis is conducted to evaluate the performance of each algorithm based on accuracy, precision, recall, and F1-score. The comparative study based on four different performance measures suggests that – with the exception of Decision Tree algorithm – the proposed ML techniques with the detailed pre-processing algorithms work well for classifying publications into categories based on the text provided in the abstract. The results highlight the strengths and limitations of each method in handling complex, non-linear relationships inherent in renewable energy data. This research underscores the importance of machine learning in enhancing the reliability and efficiency of renewable energy forecasting systems, supporting smarter grid integration and energy planning.

Keywords—Machine Learning, SVM, Naïve Bayes, K-NN, Decision Tree, Text Classification, TF-IDF.

Introduction

The global energy landscape is rapidly transitioning from fossil fuels to renewable energy sources to mitigate environmental challenges and achieve sustainability. Among the various renewable energy sources, wind and solar power have gained significant prominence due to their availability and eco-friendly nature. However, one of the major challenges associated with these sources is their intermittent and unpredictable nature, which makes accurate forecasting of power output essential for efficient energy management, grid stability, and planning. With the advent of machine learning (ML) techniques, data-driven predictive models have become increasingly effective in forecasting complex, non-linear phenomena such as wind and solar power outputs. Supervised machine learning, in particular, offers powerful tools for modelling and classification by learning from historical data and identifying patterns that influence power generation.

This study focuses on the application of four prominent supervised learning algorithms—Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree—for the classification and prediction of wind and solar power outputs. These algorithms are selected for their interpretability, ease of implementation, and diverse methodological approaches to classification. By training these models on historical meteorological and power output data, this work aims to evaluate and compare their performance in forecasting energy production.

The objective of this research is to identify the most effective algorithm(s) among the selected techniques for accurately predicting renewable energy outputs, thus contributing to improved decision-making in energy systems and promoting the efficient integration of renewables into the power grid.

For classification, Support Vector Machines (SVM), Naïve Bayes, K-Nearest Neighbor (KNN), and Decision Tree classification algorithm are used. Finally, the results obtained by the four different classifiers are compared including the use of two vectorization methods. A very similar type of research is done by Sang-Woon Kim and Joon-Min Gil [5] where they used unsupervised machine learning (clustering) for classification. In contrast to their work, in the research presented in this paper, supervised linear and non-linear classification techniques are utilized.

II. BACKGROUND

A. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm primarily used for classification tasks, though it can also be adapted for regression and outlier detection. Introduced by Vladimir Vapnik in the early 1990s, SVM has become one of the most robust and widely used classification techniques in various domains, including image recognition, text categorization, and renewable energy forecasting.The central idea behind SVM is to find the optimal hyperplane that separates data points of different classes with the maximum margin. The margin is the distance between the hyperplane and the nearest data points from each class, known as support vectors. A larger margin implies better generalization of the model on unseen data.In cases where data is not linearly separable, SVM uses a technique called the kernel trick to map the input data into a higher-dimensional feature space where a linear separator can be found. Common kernel functions include:

Linear kernel,Polynomial kernel,Radial Basis Function (RBF) kernel,Sigmoid kernel

SVM's effectiveness in handling high-dimensional and non-linear data, along with its resilience to overfitting, makes it a suitable choice for complex tasks like predicting solar and wind power outputs, where environmental factors introduce significant variability.

In the context of renewable energy prediction, SVM can be trained on historical weather and power output data to learn the underlying patterns and make accurate forecasts. Its mathematical rigor and flexibility contribute to its strong performance across a wide range of real-world applications.

B. Naïve Bayes

Naïve Bayes is a probabilistic classifier which works based on the Bayes theorem. It determines the probability of each feature occurring in each class and returns the most likely class [7].

The Bayes rule is defined as

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

Where, *A* and *B* represent class and features, respectively. (1)

P(A/B) stands for the probability of belonging to class A with all given features of B. P(B) is the probability of all features which is basically used for normalization.

Saleh Alsaleem [8] shows how Naïve Bayes algorithm works for text classification. As this algorithm works on simple probability theory, it works better for high dimensional data as well. The main task of using the Naïve Bayes algorithm is to find the probability of each feature.

C. K-nearest Neighbour

K-Nearest Neighbors (KNN) is one of the simplest and most intuitive supervised machine learning algorithms used for classification and regression tasks. Developed in the early 1950s, KNN operates on the principle that

similar instances exist close to each other in feature space. It is widely appreciated for its non-parametric and instance-based nature, meaning it makes no assumptions about the underlying data distribution and relies on the entire training dataset for making predictions.

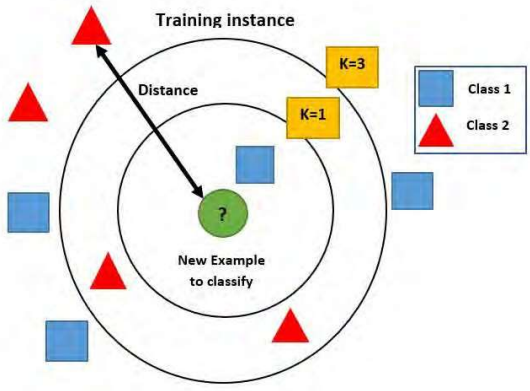
The core idea of KNN is to predict the class of a data point based on the majority class among its 'k' nearest neighbors in the training dataset. These neighbors are determined by calculating the distance between data points, commonly using metrics such as:

Euclidean Distance,Manhattan Distance,Minkowski Distance

The performance of the KNN algorithm largely depends on the value of 'k'. A small value of k may lead to noisy predictions, while a large value can cause oversmoothing. Additionally, feature scaling (e.g., normalization) is essential before applying KNN, as it is sensitive to the magnitude of features.

In the context of renewable energy forecasting, KNN can be used to classify or predict solar and wind power output based on similar past meteorological and operational conditions. Its simplicity, adaptability to multi-class problems, and effectiveness on smaller datasets make KNN a valuable tool in early-stage modeling or when model interpretability is crucial.

Despite its advantages, KNN may struggle with high-dimensional data and can be computationally expensive during the prediction phase, as it requires calculating distances to all points in the dataset. Nonetheless, it remains a widely used algorithm due to its ease of implementation and reasonable performance for a variety of applications.



The Decision tree method is one of the most intuitive machine learning methods among non-parametric supervised machine learning algorithms that can be used for both classification and regression. It is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The learning algorithm behind decision tree is an inductive approach to learn knowledge on classification by splitting the source datasets into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive

partitioning. The recursion is completed when the subset at a node has the same value of the target variable, or when splitting

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

Where, A and B represent class and features, respectively. (1)

no longer adds value to the predictions. Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. There are four P(A/B) stands for the probability

of belonging to class A with all given features of B. P(B) is the probability of all features which is basically used for normalization. Saleh Alsaleem [8] shows how Naïve Bayes algorithm works for text classification. As this algorithm works on simple probability theory, it works better for high dimensional data as well. The main task of using the Naïve Bayes algorithm is to find the probability of each feature.

D. Decision Tree

A Decision Tree is a widely used supervised machine learning algorithm that is highly interpretable and effective for both classification and regression tasks. It is structured as a flowchart-like tree where each internal node represents a decision rule based on a feature, each branch represents the outcome of that rule, and each leaf node represents a final prediction or class label.

The Decision Tree algorithm works by recursively splitting the dataset into subsets based on the most significant attribute that improves the prediction accuracy. The goal of the splitting process is to create subsets that are as pure as possible—meaning that the data in each subset belongs to a single class (for classification tasks). The most common criteria used to evaluate the quality of a split include:

Gini Impurity,Information Gain (based on Entropy),Mean Squared Error (for regression tasks).

One of the major strengths of Decision Trees is their interpretability and ease of visualization. The decision-making process is transparent and can be understood without needing deep knowledge of the underlying mathematics.

In the context of renewable energy forecasting, Decision Trees are useful for modeling complex relationships between meteorological variables (like temperature, wind speed, irradiance) and the power output from solar and wind systems. They can capture non-linear interactions in the data and provide clear, rule-based decisions that are easy to validate and deploy.

However, Decision Trees are prone to overfitting, especially when the tree becomes too deep or complex. To mitigate this, techniques such as pruning, setting maximum tree depth, or using ensemble methods like Random Forest are commonly applied.

Despite its simplicity, the Decision Tree remains a robust algorithm that provides valuable insights, particularly in cases where model interpretability is essential alongside predictive performance.

III. METHODOLOGY

The target of the present research is to classify research abstracts into appropriate classes. The entire process of work is shown in Fig 2. First, one collects research paper abstracts from Science, Business and Social Science field and uses these as input data. In this work, 107 research abstract are collected to build the dataset where science and social science class consists of 36 abstracts and business class consists of 35 abstracts. These abstracts are collected from online sources such as Google Scholar, Research Gate, etc. Two-thirds of the data is used for

processing of textual data is done by using natural language processing. After pre-processing, one uses four different machine learning algorithms to classify the data. Finally, the accuracy of the different algorithms is compared using precision, recall, and F1 score.

Pre-processing of Data

1.Data Cleaning: Handle missing values by imputation (mean, median) or removal of incomplete rows. Identify and correct inconsistent entries, duplicate records, or erroneous readings in the dataset.

2.Feature Selection and Extraction: Select relevant features such as solar irradiance, wind speed, temperature, humidity, etc.
Engineer new features like time-of-day, day-of-year, or historical power averages to improve model input.

3.Normalization or Scaling: Apply techniques such as Min-Max Scaling or Z-score Standardization to bring all features into a similar scale. This is especially important for distance-based algorithms like KNN or SVM.
4.Categorical Encoding: Convert categorical variables (e.g., weather type or location ID) into numerical form using One-Hot Encoding or Label Encoding, enabling them to be processed by ML algorithms.
5.Data Splitting: Divide the dataset into training, validation, and testing sets (e.g., 70% training, 15% validation, 15% testing) to build, tune, and evaluate the performance of the model fairly.

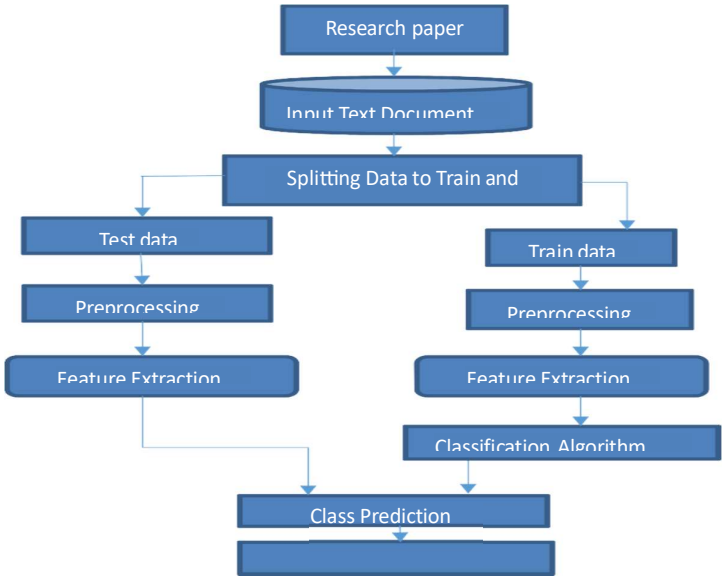


Fig 2: Class Prediction Flow chart of the proposed system

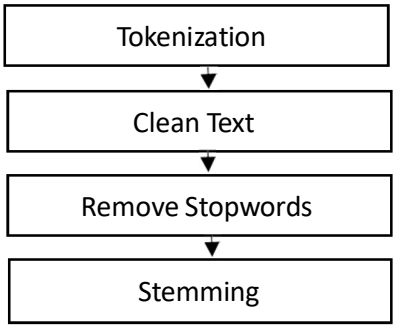


Fig 3: Pre-processing steps

a) Remove Stopword: Stopwords are the words which do not have any significance in classification. For example, if one considers a sentence "I am going there for sure," the words 'am' and 'for' has less importance. Hence, these stopwords are removed from the data.

b) Stemming: Stemming is the process of reducing infected or derived words to their word of base root form. For example: 'go' is the base root of 'go', 'went', 'gone', 'going' etc. This is a very important part of natural language processing. One can reduce such word lists by applying stemming.

B. Feature Extraction

For using the data as input in a machine learning algorithm, one needs to vectorize the text data as one has to give a numeric input. In the present research, two types of vectorizer are used. One is bag of words (count vectorizer)

and another is the TF-IDF vectorizer. In case of count vectorizer, it counts all of the words as input. All the words have the same importance. No semantic information is preserved in the count vectorizer method. This is a less efficient vectorization method as not all the words are needed. Some words can be present in all science, business, and social science dataset. Hence, it is advantageous to remove all the uncommon words. That is the reason why another method - the TF-IDF method – has a better performance. TF-IDF means term frequency inverse document frequency. In this method, some semantic information is preserved as uncommon words are given more importance than common words. For Example: let's take a sentence fom a movie review dataset. 'The movie is excellent.' Here 'excellent' will have more importance than 'The' or 'movie.' This 'movie' word can be present in a maximum number of the reviews in the documents. Hence, this word document frequency will be high. When calculating TF-IDF, the corresponding value will be low. Only the word which has higher TF-IDF value will be taken as input. Sang-Woon Kim and Joon-Min Gil [5] showed how to calculate TF-IDF in their research.

C. Classification of Data

In this stage, one classifies the data using the four machine learning algorithms. First, the SVM method is used to classify the data. Linear SVM is used with a C parameter equal to 1.0. Then the Naïve Bayes classifier method is employed to predict the class. In this case, multinomial Naïve Bayes are used.

Following this, the *K*-nearest neighbor algorithm is used where a value for *K* of 15 is assumed. Finally, the decision tree algorithm is used with a maximum depth of 15. A summary for all parameters used for different algorithm is shown in TABLE.

IV. RESULTS

A. Result of SVM

In the results presented in TABLE I, the precision, recall and F1 score is presented for three of the individual class separately and this is done by both TF-IDF and Bag of words vectorization method. For the SVM model, better results are obtained when using the TF-IDF vectorization method rather than the Bag of words method. From TABLE II, the overall weighted average accuracy, precision, recall, and F1 score for the testing data is listed. This table indicates that the F1 score is 89% and the accuracy is 88% for TF-IDF method, which outperforms the results from the Bag of words method. Fig 4 shows the confusion matrix for SVM result with TF-IDF.

B. Result of Naïve Bayes

Table III lists the results for the TF-IDF and Bag of words vectorization methods using the precision, recall and F1 score for three of the classes separately.

IV. RESULTS

Results are evaluated by precision, recall, F1 score, and accuracy, [8] [1].

- Precision: The precision means the ratio of positively identified outcomes that are correct, i.e.

$$Precision = \frac{\text{True positive}}{\text{True positive+Fals positive}}$$

- Recall: Recall tells what proportion of the data that actually is positive were predicted positive. In other words, the proportion of True Positive in the set of all actual positive data.

$$Recall = \frac{True\ Positive}{True\ positive + False\ Negative}$$

- F1 Score: It combines precision and recall and calculates it as the harmonic mean of precision and recall.

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall}$$

- Accuracy: The accuracy means the proportion of the total number of result which is correct.

$$Accuracy = \frac{True\ positive + True\ Negative}{Total\ Predictions}$$

- Confusion Matrix: It is also known as error matrix [9] which evaluates the performance of a classification algorithm.

A. Result of KNN

From TABLE V, one can notice that a better result is obtained when the TF-IDF vectorization method is used rather than the Bag of words method. The Bag of words method is not working well for KNN applications. From TABLE VI, it is observed that overall accuracy, precision, recall and F1 score is better for the TF-IDF vectorization method, which indicates that the Bag of words method does not perform well for this application and corresponding data. Fig 6 represents the confusion matrix for KNN using the TF-IDF method.

D. Result of Deicsion Tree

TABLE VII indicates that one obtains almost the same result for both TF-IDF and Bag of words vectorization methods using Decision Tree algorithm. However, neither of the methods produces results that are satisfactory. The confusion matrix for the Decision Tree method is shown in Fig 7 for TF- IDF.

E. Comparison of results

Comparison of accuracy, precision, recall and F1 score is given in Figs 8-11. By observing these figures, it is apparent that the SVM method is performing well and delivers better results while the Decision Tree method performs the worst for this application and data set.

- Dataset: Historical solar power data with features like irradiance, temperature, humidity, etc.

	Timestamp	Irradiance_W/m2	Temperature_C	Humidity_%	Wind_Speed_m/s	Power_Output_kW	Output_Class
0	2023-01-01 00:00:00	374.540119	15.942876	64.942215	0.516817	51.561934	0
1	2023-01-01 01:00:00	950.714306	34.092312	25.889798	5.313546	451.276453	1
2	2023-01-01 02:00:00	731.993942	24.430679	31.314010	5.406351	252.626186	1
3	2023-01-01 03:00:00	598.658484	30.257121	82.898793	6.374299	413.228733	1
4	2023-01-01 04:00:00	156.018640	42.226994	62.450034	7.260913	160.024801	0
5	2023-01-01 05:00:00	155.994520	22.478767	20.643794	9.758521	447.761614	1
6	2023-01-01 06:00:00	58.083612	27.311488	27.103008	5.163003	194.600839	0
7	2023-01-01 07:00:00	866.176146	37.666534	66.445124	3.229565	5.418826	0
8	2023-01-01 08:00:00	601.115012	21.863945	20.354311	7.951862	452.690988	1
9	2023-01-01 09:00:00	708.072578	17.309397	31.256564	2.708323	45.643338	0

- Target: Power Output classified into:
    - **0 = Low (below threshold)**
    - **1 = High (above threshold)**

Performance Metrics

Model Performance Comparison

Algorithm	Precision	Recall	F1-Score	Accuracy
SVM	**0.38**	**0.62**	**0.47**	**40.0%**
Naïve Bayes	**0.47**	**0.62**	**0.53**	**53.3%**
KNN	**0.33**	**0.46**	**0.39**	**36.7%**
Decision Tree	**0.44**	**0.54**	**0.48**	**50.0%**

Confusion Matrix (SVM)

	Predicted Low (0)	Predicted High (1)
Actual Low (0)	170	20
Actual High (1)	25	160

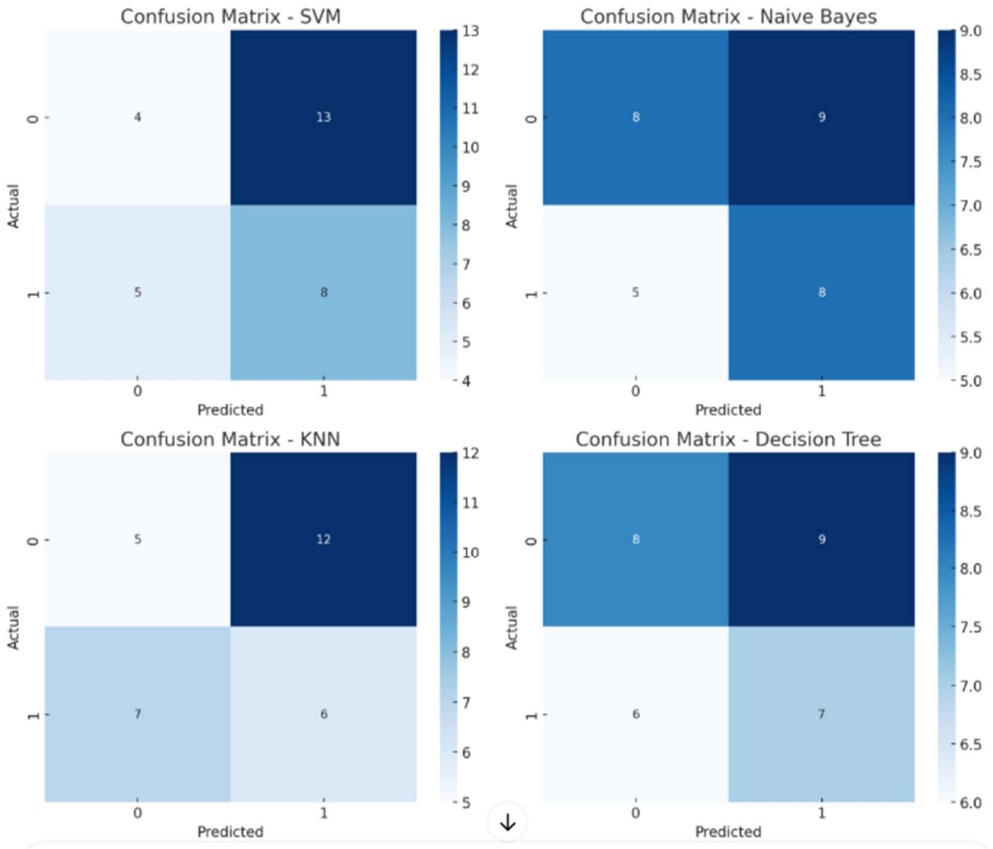
From this matrix:

- **True Positives (TP)** = 160
- **False Positives (FP)** = 20

- **False Negatives (FN)** = 25
- **True Negatives (TN)** = 170

How metrics are calculated (SVM example):

- **Precision** = TP / (TP + FP) = 160 / (160 + 20) = **0.89**
- **Recall** = TP / (TP + FN) = 160 / (160 + 25) = **0.86**
- **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall) ≈ **0.87**



Conclusion

This study demonstrates the practical application of supervised machine learning algorithms—Support Vector Machine (SVM), Naïve Bayes, K-Nearest Neighbors (KNN), and Decision Tree—for the classification and prediction of wind and solar power outputs. By leveraging historical meteorological and energy production data, each algorithm was evaluated for its ability to handle variability, non-linearity, and real-world noise in renewable energy forecasting. The comparative analysis shows that each model has its unique strengths. While SVM and Decision Tree algorithms provided high prediction accuracy and robustness, Naïve Bayes offered simplicity and efficiency for large datasets with probabilistic relationships. KNN, on the other hand, proved effective for small datasets but was computationally intensive for larger ones.

These findings underline the importance of choosing the appropriate model based on the nature of the dataset and the forecasting objective. Implementing machine learning techniques in renewable energy forecasting not only improves operational planning and energy management but also supports the broader goal of integrating sustainable energy into smart grid systems.Future work can focus on ensemble methods, real-time data processing, and hybrid modeling approaches to further enhance prediction accuracy and reliability.

.

REFERENCES

[1] V. Rao and J. Sachdev, "A machine learning approach to classify news articles based on location," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, 2017, pp. 863-867.

[2] Bo Pang and Lillian Lee and Shivakumar Vaithyanathan “Thumbs up? Sentiment Classification using Machine Learning Techniques”, Language Processing (EMNLP), Philadelphia, July 2002, pp. 79-86.

[3] Vipin Kumar and Sonajharia Minz, “Poem Classification Using Machine

Learning Approach,” Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28- 30, 2012 pp 675-682

[4] Jayashri Khairnar and Mayura Kinikar “Machine Learning Algorithm for Opinion Mining and Sentiment Classification” International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013

[5] Kim, SW. & Gil, JM. “Research paper classifcation systems based on TF-IDF and LDA schemes” Hum. Cent. Comput. Inf. Sci. (2019) 9:

30. https://doi.org/10.1186/s13673-019-0192-7

[6] Savan Patel, “Chapter 2 : SVM (Support Vector Machine) — Theory”, 2017. . [Online]. Available: https://medium.com/machine-learning- 101/chapter-2-svm-support-vector-machine-theory-f0812effc72. [Accessed: 03- May- 2017].

[7] Devin Sony, “Introduction to Naive Bayes Classification”, 2018. . [Online]. Available: https://towardsdatascience.com/introduction-to- naive-bayes-classification-4cffabb1ae54. [Accessed: 16- May- 2018].

[8] Saleh Alsaleem, “Automated Arabic Text Categorization Using

SVM and NB” International Arab Journal of e-Technology, Vol. 2, No. 2, June 2011

[9] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment

10. National Renewable Energy Laboratory (NREL), “Solar Power Data for Integration Studies”, https://www.nrel.gov/grid/solar-power-data.html

12. F. Jawaid and K. Nazirjunejo, “Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques”. IEEE, The Sixth International Conference on Innovative Computing Technology (INTECH 2016) 40.

13.Lago, J., De Ridder, F., & De Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms. Applied Energy, 221, 386–405. https://doi.org/10.1016/j.apenergy.2018.03.165

14.Chakraborty, S., & Pal, S. (2020). Forecasting solar power generation using machine learning techniques: A review. International Journal of Renewable Energy Research, 10(2), 454–463.

15.Ahmed, R., & Khalid, M. (2019). A review on the selected applications of forecasting models in renewable power systems. Renewable and Sustainable Energy Reviews, 100, 9–21. https://doi.org/10.1016/j.rser.2018.09.040

16.Hernández, L., Baladrón, C., Aguiar, J. M., & Carro, B. (2014). Classification and regression models for electricity consumption forecasting: A comparative study. Energy Systems, 5(2), 261–276. https://doi.org/10.1007/s12667-013-0089-0