iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.48056

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

SenseWorth – A Tweets Classifier

Priyanshu T Agarkar[1], Pranav Chopdekar[2], Sahil Gujar[3], Komal Chitnis[4]
[1, 2, 3, 4]*Department of Computer Engineering, A.P Shah Institute of Technology, Thane(M.H), India 400615*

Abstract: ***A lot of people use tweeter to provide their opinions on various topics including sports, politics, finance, etc. Now the information that generally is present on Twitter does not have any means to check whether the data being twitter is correct or not. So in order to check the authenticity of the data,the data must be classified into true and false. Firstly, the dataset would be created using python by devising the code for the same. The code thus designed would be able to extract the tweets with the amount specified by the user. Hence the CSV would be created. After the creation of the CSV, data pre-processing would be applied to the data such that all the unnecessary data such as emojis, words, and information would be removed automatically, and thus we would receive the tweet without any kind of stop-words. This cleaned data would further be tested and trained with the help of machine learning algorithms. These Machine learning algorithms would generally be used to classify the data according to the domain and provide the user with an authenticated answer whether the tweet is true or not along with its accuracy. This helps the user to identify the tweet and thus provide an authenticated answer on which informationto believe and to which we should not.***

Keywords**:** ***Data Extraction, Data Pre-processing, Support Vector Machine, Random Forest, Multinomial Naïve Bayes***

I. INTRODUCTION

Classification of tweets is the task of finding the opinions and information that people post about specific topics of interest. Be it a political topic or a sports topic, the opinions of people matter, andit affects the decision-making process of people. The first thing a person does when he or she wantsto write a particular tweet is to see the kind of reviews and opinions that people have written about a particular post on Twitter Social media such as Facebook, blogs, and Twitter have become a placewhere people post their opinions on certain topics. The classification of the tweets of a particular subject has multiple uses, including classifying the tweets in various domains like politics, sports, health, and so on. Classification of tweets can be categorized into two categories that are true and false. The two types of classification categories in this classification experiment are real and fake tweets. The data, being labeled by humans, has a lot of noise, and it is hard to achieve good accuracy. Currently, the best results are obtained by the Support vector machine(SVM) for a feature set containing stemming that gives an accuracy of 82.55. The main algorithms used in this project are SVM and Multinomial Naive Bayes and we would be comparing these in the upcoming sections. The report is organized in the following way. The four main important topics include data pre- processing, the machine learning algorithms used, the tools required to execute the project as well as the feature extraction techniques along with features used.

Firstly, we will need to extract the data from various sources that are available on Twitter. Then thedata being extracted would be classified into four domains namely politics, sports, health, and finance. The tweet can directly be searched by giving the URL or directly the tweet in the search bar. After we click on the search bar the result containing the classification of the tweet and its accuracy would be displayed on the screen. This classification and accuracy are generally presented by using machine learning algorithms like multinomial naïve Bayes, Decision tree, random forest, etc.

The tweets that are extracted are extracted using python. The modules being used are Snscrape andTweepy. We need the consumer key and the access key that is that will be acquired with the help ofa developer account. The developer account provides a path to extract tweets related to a specific domain.

II. LITERATURE SURVEY

In this [1] work they proposed a novel text analysis based computational approach to automaticallydetect fake news. The results obtained for a test dataset show promise in this research direction. Forexploratory purposes, they have created another dataset of 345 “valid” news articles. This dataset includes an equal number of news reports from three well-known and largely respected news agencies: National Public Radio, New York Times, and Public Broadcasting Corporation. They created a sample of 345 valid news articles as above and complemented it with another set of 345 randomly selected articles from the Kaggle dataset on “BS” news. They used LIWC (Linguistic Analysis and Word Count) package to obtain linguistic features for each of the articles. Each featurewas normalized using Z-score normalization.

They undertook an 80-20 split on the data for the training and test sets. They created multiple machine learning models based on well-established algorithms such as logistic regression, support vector machine, random forest, decision tree, k- neighbors classifier etc. and focused on the performance of the algorithms for the test set. Among these algorithms, Support Vector Machine method gave the best prediction results. The results obtained for a test dataset show promise in this research direction. Dataset used was Kaggle Fake News, 345 "valid" news articles and algorithms used were logistic regression, support vector machine, random forest, decision tree,k- neighbors classifier. Their model gave accuracy of 86%.

Junaed, Tawkat, Anindya and Sadia in [2] paper, has benchmarked three different datasets, of which the largest and most diverse one was created by them, to evaluate the performance of various suitable techniques. According to the research, they also used several cutting-edge deep learning models, and the results have been encouraging. The authors of this study provided a comprehensive performance analysis of various techniques on three separate datasets. On a dataset with fewer than 100k news articles, they demonstrated that Naive Bayes with n-gram can get results that are comparable to those of neural network-based models. The length of the dataset and the details provided in a news story have a significant impact on how effectively LSTM-based models perform. It is more likely for LSTM-based models to avoid overfitting. Furthermore, cutting-edge models like C-LSTM, Conv-HAN, and character level C-LSTM have demonstrated significant promise, calling for continued focus on these models in the detection of fake news. Finally, they conduct a topic-based study that highlights how challenging it is to accurately identify false news in the fields of politics, health, and research.

In this [3] paper they have presented various methods which can be used for finding results and comparing them with which we can understand which method is better and it gives a max correct information about fake or real. The dataset used was Chile earthquake 2010. The accuracy of 62.47% for Logistic Regression, 84.56% for Naïve Bayes and 89.34% for SVM ; TF-IDF, we got 69.47% for Logistic Regression, 89.06% for Naïve Bayes and 89.34% for SVM.

In this [4] research paper, firstly, they have investigated the existing models for fake news detection using content and context-based information. After an initial investigation, they have performed extensive experiments using a multi-class dataset (FNC-based fake news dataset) and employed different machine learning algorithms. In this exploration, they have found that among the different machine learning algorithms used, Gradient Boosting with optimized parameters performs the best for a multi-class fake news dataset. In the existing research, benchmark results are available based on two classes dataset, classifying news as fake or real. Work on multi-class prediction is limited. In this work, they have implemented various machine-learning models using a multi-class dataset for the fake news classification and achieved an accuracy of 86% with the gradient boosting model. This demonstrates its suitability for a multi-class textual classification problem. Apart from machine learning based models using content-level features, a hybrid model can be more helpful to classify news using both the content and context-based information of the news.

The study[5] concludes that despite the fact that bogus news and posts can clearly be detected using a variety of machine learning techniques, however, the characteristics and features of fake news on social media networks are constantly changing, making it difficult to categorise. However, computing hierarchical features is the primary distinguishing trait of deep learning. Numerous research projects are utilising deep learning techniques in a variety of applications, including audio and speech processing, natural language processing and modelling, information retrieval, objective recognition and computer vision, as well as multimodal and multi-task learning. These techniques include convolutional neural networks, deep Boltzmann machines, deep neural networks, and deepautoencoder models.

In [6], by explaining how to visualise the accuracy exams and taking into account automating fashioned news distinguishing evidence in Twitter datasets, the authors of this research suggested a strategy for understanding manufactured news messages from Twitter tweets. Then, to demonstrate the success of the grouping execution on the dataset, they separately performed a correlation between five well-known Machine Learning calculations, including the Support Vector Machine, Naive Bayes Method, Logistic Regression, and Recurrent Neural Network models. The results of their exploratory study showed that SVM and Nave Bayes classifier outperformed various

In this [7] study, the authors used the publicly accessible LIAR dataset, which wass gathered from POLITIFACT.COM for the detection of fake news and includes links to each case's supporting documentation. On this dataset, the accuracy in all previous works had been approximately 30%. In this study, they make use of model ensemble approaches to improve the LIAR dataset's ability to identify bogus news. Additionally, they attempted to reduce the complexity of the problem statement to a binary classification and used the same ensemble approaches to develop a more realistic strategy for precise calculation. Here Bagging Classifier and AdaBoost attain 70% accuracy in precision, F1-Score and recall. These are confirmed by running the model for 150 iterations.

In this [8] paper, the project offers three methods for using NLP to identify fake news—that is, newsitems that are deceptive and originate from unreliable sources. To extract text data, they suggest utilising the Python sci-kit-learn module, which has utilities like the Count Vectorizer, Tf-IDF Vectorizer, and Hashing Vectorizer. The training model was then been run, classified using user input, and evaluated for accuracy and precision using the findings of the confusion matrix. The fivemodels' results are examined, and a combination of machine learning and natural language processing methods is selected. These five models were produced by mixing various NLP strategieswith all machine learning algorithms. A machine learning model that has been trained is connectedto a user interface that has been created. To forecast user-inputted news, a passive aggressive classifier with a TF-IDF vectorizer was trained and employed.

Uma, Siddarth and Shankar in [9] paper, has explained that they used machine learning, natural language processing, and artificial intelligence ideas to perform binary classification on a variety of online news articles. They sought to give users the option of classifying news as true or phoney andchecking the legitimacy of the website disseminating it. The system, which was developed in threesections, is explained in this essay. The first section uses a machine learning classifier and is static.They researched the model and trained it using four different classifiers before selecting the most effective one to use in the end. The second component, which is dynamic, uses the user's term or text to search internet for the actual likelihood of the news. The final component confirms the legitimacy of the user-provided URL. They used Python and its Sci-kit libraries. With an accuracyof 65%, Logistic Regression proved to be the most accurate model. In order to improve the performance of logistic regression, they employed grid search parameter optimization, which provided us a 75% accuracy. In their model the user can research news articles, keywords, and the legitimacy of websites online. The dynamic system's accuracy is 93%, and it gets better with each repetition.

Prasad, Suyash, Rhucha, Prashant and Sumitra in [10] paper, offer a solution by outlining a machine learning-based fake news detection technique. Prerequisite information acquired from different news websites is needed for this model. Data is extracted from websites using a technique called web scraping, which is then used to produce databases. The real dataset and the false dataset are the two main categories into which the data is divided. Classifiers like Random Forest, Logistic Regression, Decision Tree, KNN, and Gradient Booster are used to categorise data. The data is categorised as either true or false based on the output that was received. Based on it, the user can determine whether the news being provided on the web server is real or fraudulent. In this research,they investigated a computerised model for confirming news retrieved from social media, which offers instructive examples for identifying fake news. Once it has been shown that even the most fundamental algorithms in fields like AI and machine learning can generate a respectable outcome on a crucial subject like the spread of false information around the world. Therefore, the results of this study point to the possibility that such systems could be very beneficial and successful in resolving this important problem.

III. PROBLEM STATEMENT

A lot of information on Twitter is not classified whether it is true or false. This information is tweeted on Twitter by various users all over the world. This information has no authentication and thus can be misleading. It’s necessary to classify the information so that people would be able to believe only in the information that is actually true. This classification would be done using machine learning algorithms by extracting the data in python using various modules.

IV. OBJECTIVE AND SCOPE

A. Extracting data from social media using Python (Twitter)
B. Identifying the domain of the data
C. Using various classification algorithms for classifying the data
D. Finding the accuracy of the data
E. Scope mainly consists of Twitter
F. Stemming and cleaning should be done in order to increase the accuracy of the outcome.

V. EXPERIMENTAL SETUP

A. Software Requirement
1) Operating System – Windows,Linux-Ubuntu
2) Minimum 4GB Ram
3) 32 bit CPU or 64 bit CPU(Intel/AMD architecture)
4) Desktop or Laptop

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

B. Hardware Requirements

1) Intel Core i3 10th generation processor or higher
2) RAM – 4 GB or above
3) ROM – 128 MB or above

VI. METHODOLOGY

1) *Identifying the Process:* Before starting any kind of project it is necessary to research the topic and get to know how each and every component of the particular topic works. Identification generally includes mapping out how the project is going to proceed and in what way the team would be able to execute/prepare theproject in the given time frame.
2) *Creation of Dataset:* As our project is generally based on classification, a dataset would be needed to be created. This dataset would be created depending on the domains the team has decided to classify. In our case, the domains consist of four domains that are sports, health, finance, and politics.The tweets would be extracted from twitter directly with the help of customer keys and access keyswhich would be acquired through a Twitter developer account. Through the help of the customer key and access key, a code would be written in python that would be able to get as many tweets as the user wants in order to create the database/dataset.
3) *Designing the Front End:* A front is a page that the user would see on its side in order to access the contents of the page. The front end would be designed in such a way that would be minimalistic and on-to-point.The front end would consist of a search bar along with an option to select any 4 of the domains depending on which the user wants to get their information classified. After clicking on the next button, the data would be classified into true or false along with it accuracy.
4) *Labelling of Dataset:* Dataset should be labelled according to the domians such that it helps to increase the accuracy of the result being displayed on the screen to user. This will also help to accurately classify the data.
5) Implementation and classification and accuracy algorithm
6) Implementation would consist of various machine-learning algorithms that would be used to classifythe data along with providing accuracy. These algorithms mainly consist of naïve Bayes, logistic regression, etc. Defining them to perform functions on the dataset is the main goal to provide the required result.

VII. PROPOSED SYSTEM

A. Flow chart

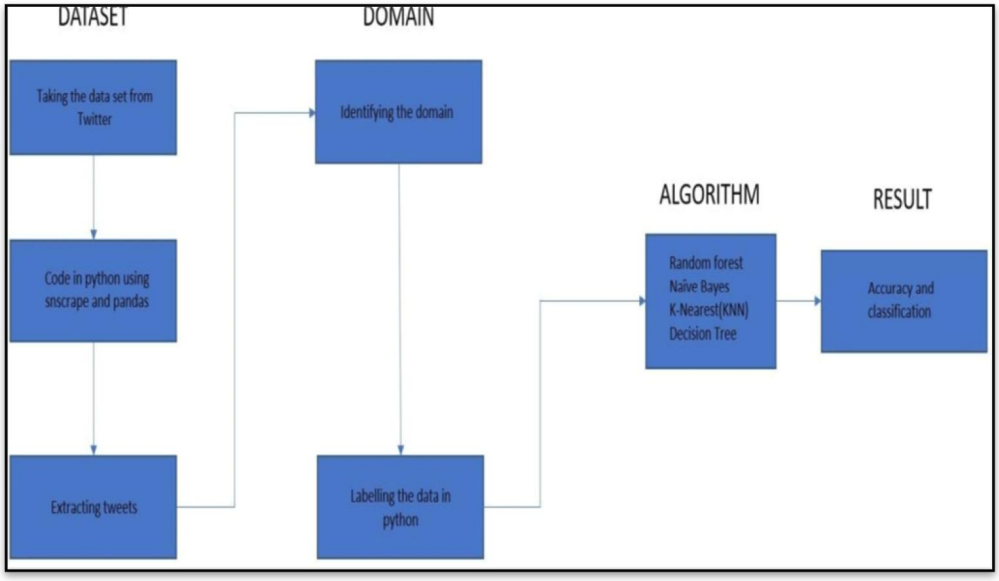


Fig 1 – Flow chart of SenseWorth application

B. Use Case Diagram

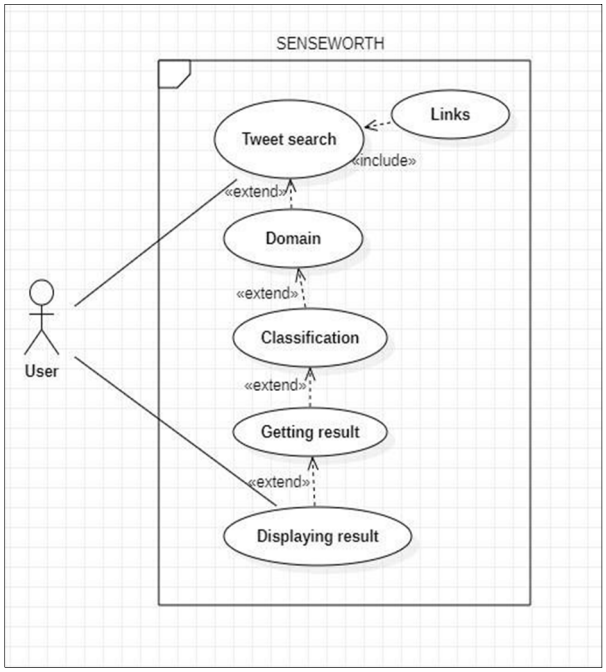


Fig 2 – Use Case Diagram of SenseWorth application

C. Data Flow Diagram (Level 0)

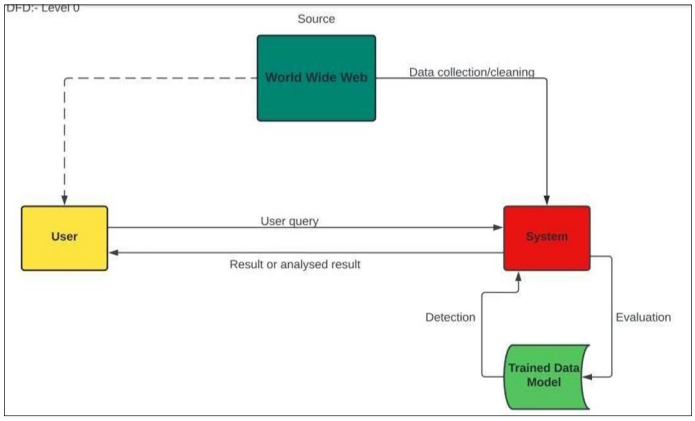


Fig 3 – DFD (Level 0) of SenseWorth application

D. Data Flow Diagram (Level 1)

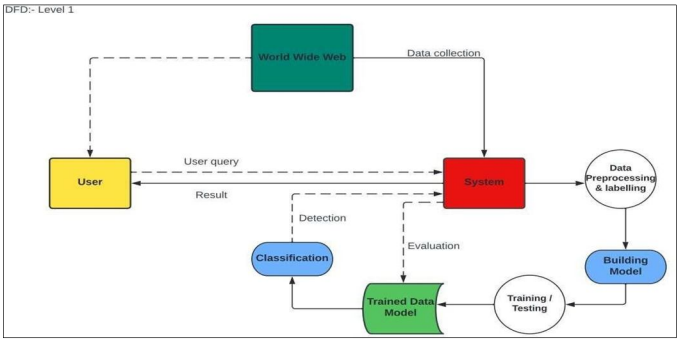


Fig 4 – DFD (Level 1) of SenseWorth application

E. Data Flow Diagram (Level 2)

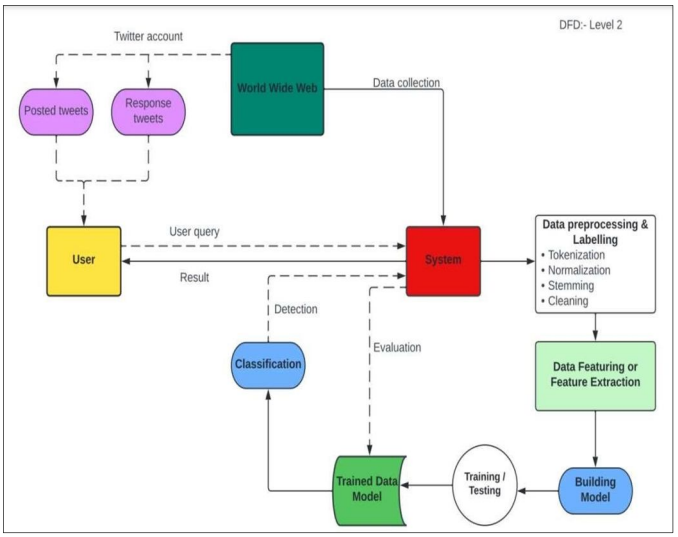


Fig 5 – DFD (Level 2) of SenseWorth application

F. Activity Diagram

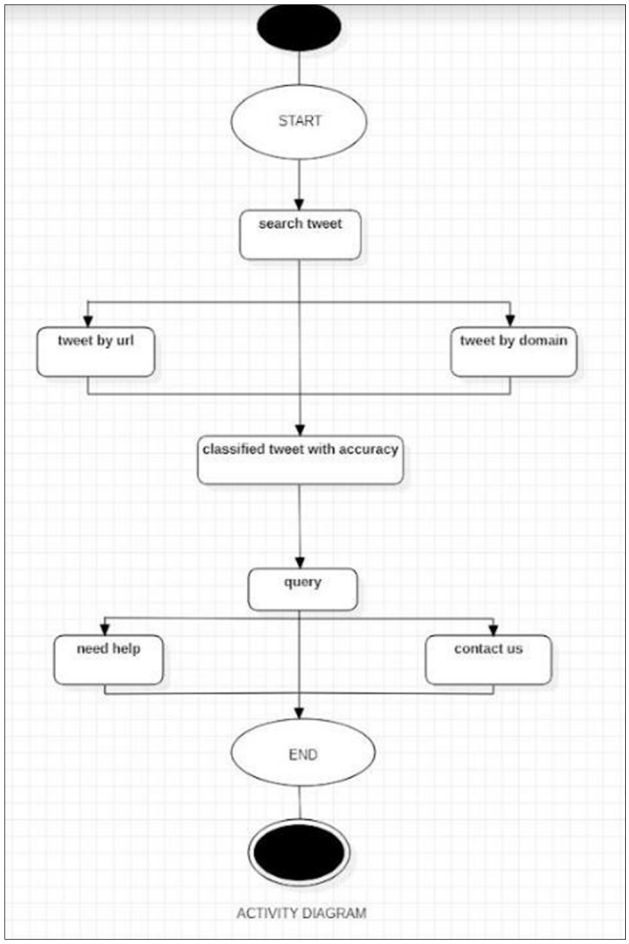


Fig 6 Activity Diagram of NewsIN application

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

VIII. PROJECT IMPLEMENTATION

A. Algorithm:

1) Import all the Modules.
2) Collect all the data from twitter with the help of those modules
3) Clean the data and find the patterns present in it
4) Then split data in training and testing
5) Train data with the help of machine learning algorithms like Multinomial Naïve Bayes, Linear Regression, Decision Tree etc
6) Then load the models in the webpage
7) Create webpages and connect them with the help of python
8) Classify the tweet with the accuracy
9) Then if any query present you can visit need help or contact us page

B. Output Screenshots

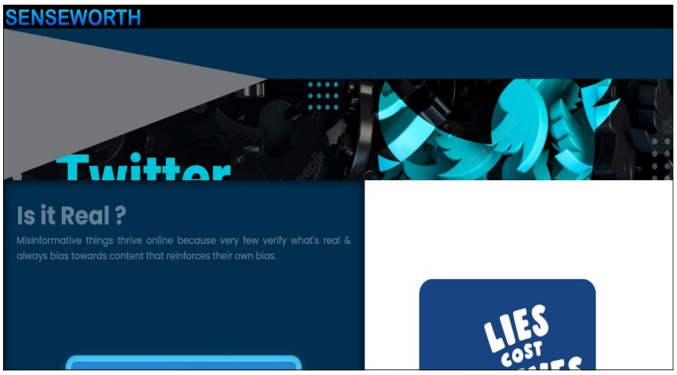


Fig 7 Main display of SenseWorth Application

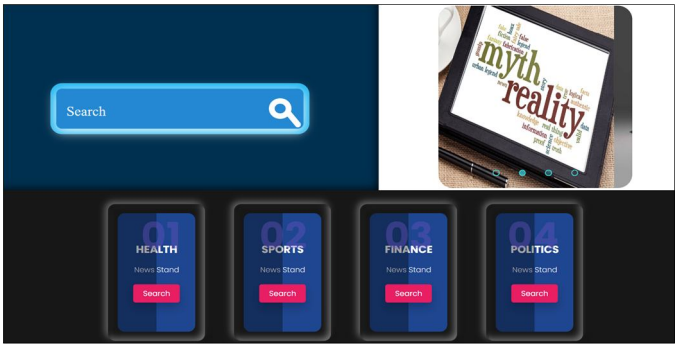


Fig 8 Domain wise display of SenseWorth Application

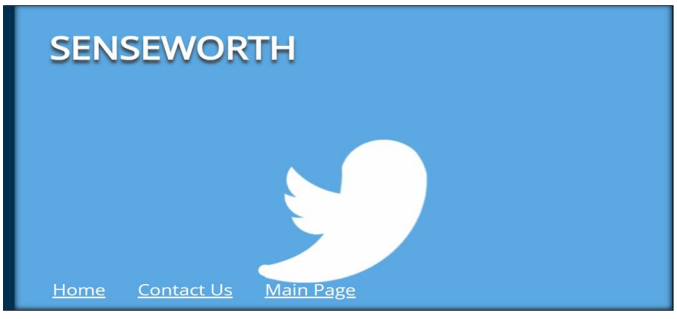
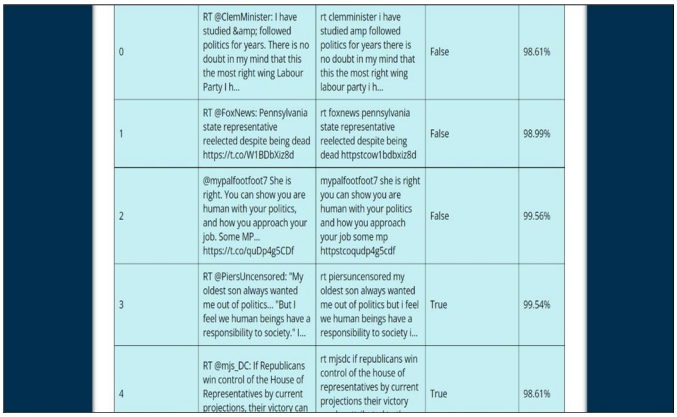


Fig 9. Result page display of SenseWorth Application



0	RT @ClemMinister: I have studied & followed politics for years. There is no doubt in my mind that this the most right wing Labour Party I h...	rt clemminister i have studied amp followed politics for years there is no doubt in my mind that this the most right wing labour party i h...	False	98.61%
1	RT @FoxNews: Pennsylvania state representative reelected despite being dead https://t.co/W1BDbXiz8d	rt foxnews pennsylvania state representative reelected despite being dead httpstcow1bdbxiz8d	False	98.99%
2	@mypalfootfoot7 She is right. You can show you are human with your politics, and how you approach your job. Some MP... https://t.co/quDp4g5CDf	mypalfootfoot7 she is right you can show you are human with your politics and how you approach your job some mp httpstcoqudp4g5cdf	False	99.56%
3	RT @PiersUncensored: "My oldest son always wanted me out of politics... "But I feel we human beings have a responsibility to society." I...	rt piersuncensored my oldest son always wanted me out of politics but i feel we human beings have a responsibility to society i...	True	99.54%
4	RT @mjs_DC: If Republicans win control of the House of Representatives by current projections, their victory can	rt mjsdc if republicans win control of the house of representatives by current projections their victory	True	98.61%

Fig 10. Display of original and clean news with prediction and accuracy in SenseWorth Application

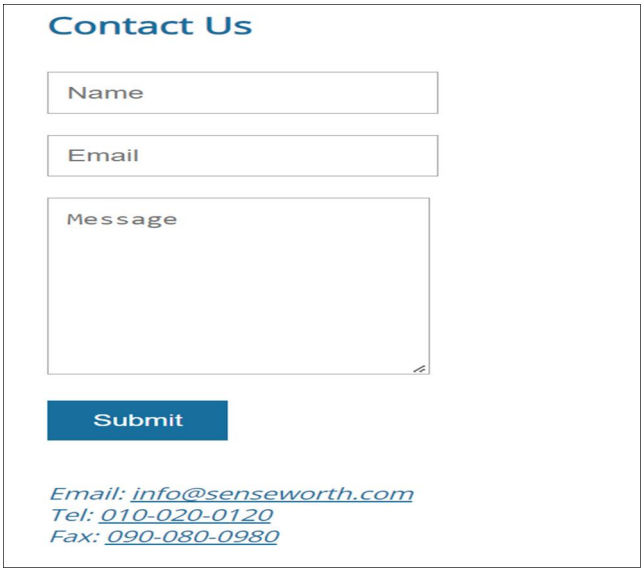


Fig 11. The Contact Us page of SenseWorth Application

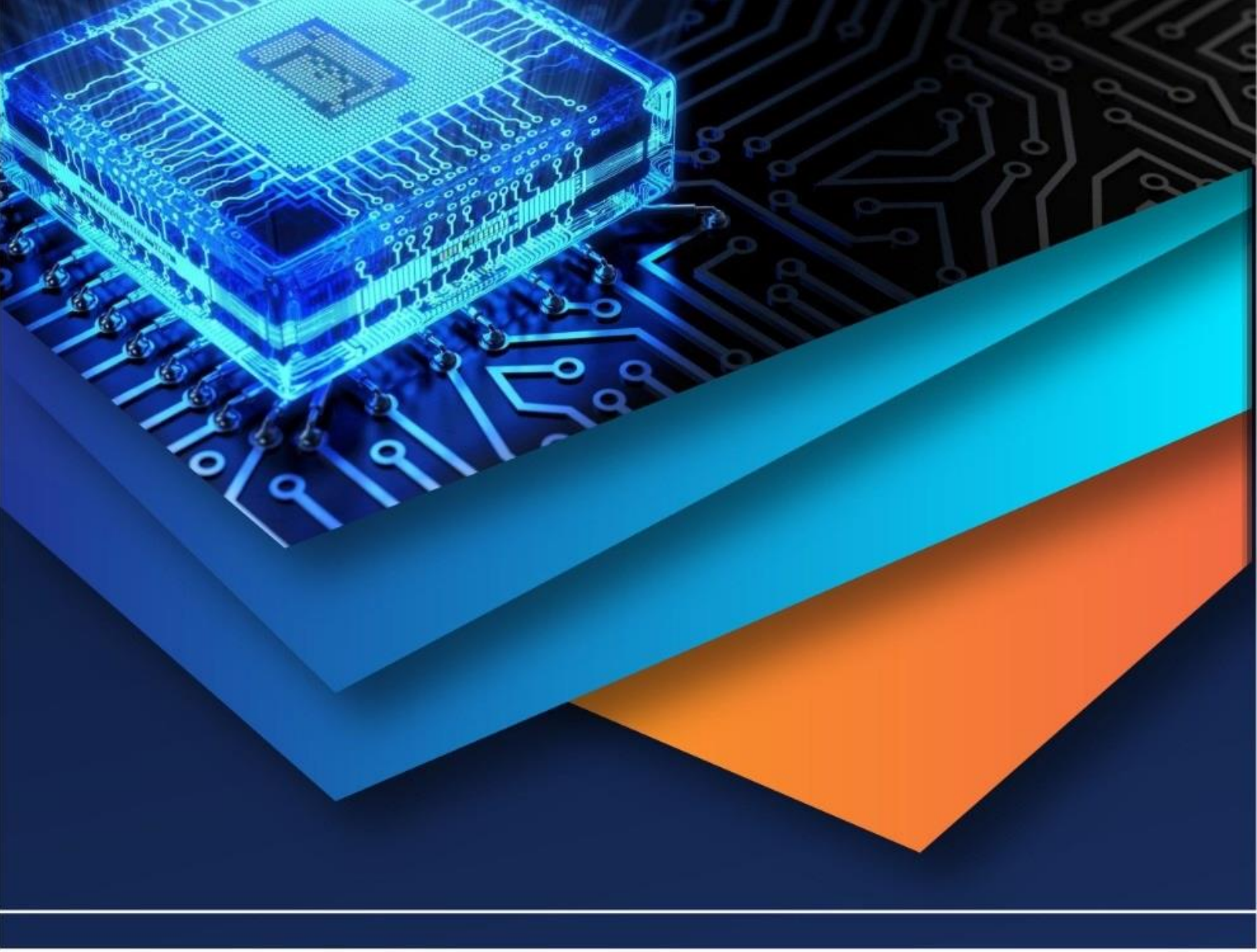
IX. CONCLUSION

In this project we were able to classify the tweets as per their domains. The main part of this project was the extraction of tweets and cleaning them to get proper insights and patterns. Then training and testing the model was another phase. Loading that model with the web page created and then connecting them to a server. Also we learned how new data can be gathered and we can fill that and match data. This provided us with the insights of the working of a classification application. This application will become the base for any future work on which anyone might want to work. Whoever will require the basics can work on it.

REFERENCES

[1] Singh, Vivek, et al. "Automated fake news detection using linguistic analysis and machine learning." International conference on social computing, behavioral-cultural modeling, & predictionand behavior representation in modeling and simulation (SBP-BRiMS). 2017
[2] J. Y. Khan, Md. T. I. Khondaker, A. Iqbal, S. Afroz, "A Benchmark Study on Machine LearningMethods for Fake News Detection", 12 May 2019.
[3] A. A. Tanvir, E. M. Mahir, S. Akhter and M. R. Huq, "Detecting Fake News using Machine Learning and Deep Learning Algorithms," 2019 7th International Conference on Smart Computing& Communications (ICSCC), 2019, doi: 10.1109/ICSCC.2019.8843612.
[4] R. K. Kaliyar, A. Goswami and P. Narang, "Multiclass Fake News Detection using Ensemble Machine Learning," 2019 IEEE 9th International Conference on Advanced Computing (IACC), 2019, pp. 103-107, doi: 10.1109/IACC48062.2019.8971579.
[5] S. I. Manzoor, Dr. J. Singla, Nikita, "Fake News Detection Using Machine Learning approaches: A systematic Review", Proceedings of the Third International Conference on Trends in Electronicsand Informatics (ICOEI 2019) , IEEE Xplore Part Number: CFP19J32-ART, ISBN: 978-1-5386- 9439-8.
[6] G. Bharath, K. J. Manikanta, G. B. Prakash, R. Sumathi and P. Chinnasamy, "Detecting Fake News Using Machine Learning Algorithms," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402470.

[7] L. Waikhom, R. S. Goswami, "Fake News Detection Using Machine Learning", International Conference on Advancements in Computing & Management (ICACM), 2019.
[8] U. Dabholkar, R. Kalapurackal, S. Timapur, Prof. Allan Lopes, "Fake News Detection Using Machine Learning", International Journal of Creative Research Thoughts (IJCRT), Volume 9, Issue5 May 2021, ISSN: 2320-2882.
[9] U. Sharma, S. Saran, S. M. Patil, "Fake News Detection using Machine Learning Algorithms",International Journal of Engineering Research and Technology (IJERT), Volume 9, Issue 3, ISSN:2278-0181, Published in 2021.
[10] P. Kulkarni, S. Karwande, R. Keskar, P. Kale, S. Iyer, "Fake News Detection using Machine Learning" ITM Web of Conferences 40, 03003 (2021), ICACC-2021, doi: https://doi.org/10.1051/itmconf/20214003003.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY