CSCI 434 FINAL PROJECT

Github Link: https://github.com/DurinMMII/coll400

TABLE OF CONTENTS

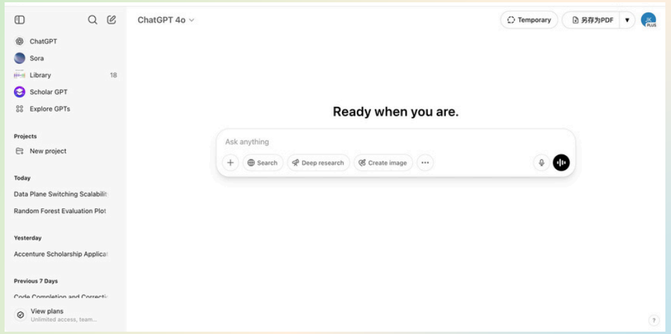
01 | PROJECT OBJECTIVE



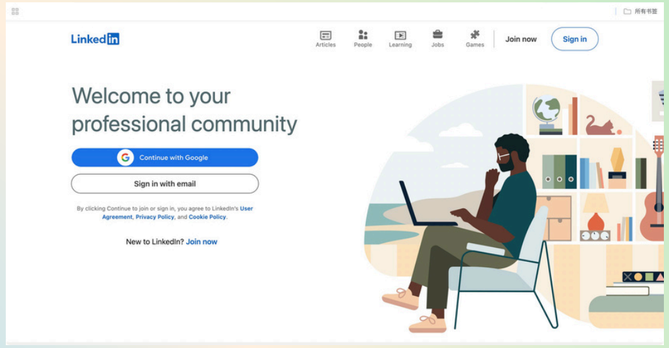

- Monitoring and analyzing computer network traffic, particularly from popular internet websites such as ChatGPT, Linkedin, Reddit, and Wikipedia.
- Extracting features from website traffic data and build models to classify the specific website from which the traffic originated.
- Comparing different models based on their performances and analyzing based on each model’s advantages and disadvantages

02 DATA COLLECTION

Visited Website

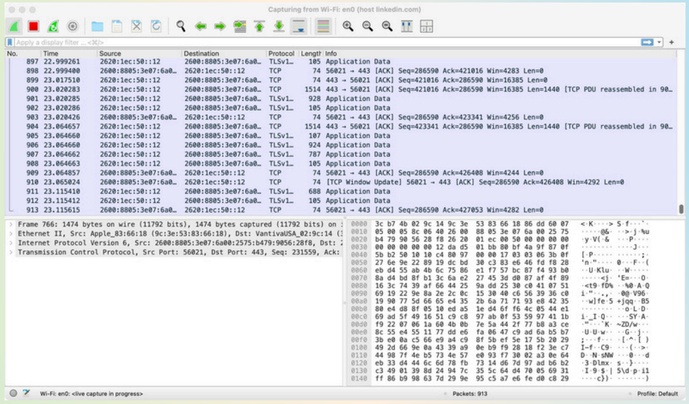


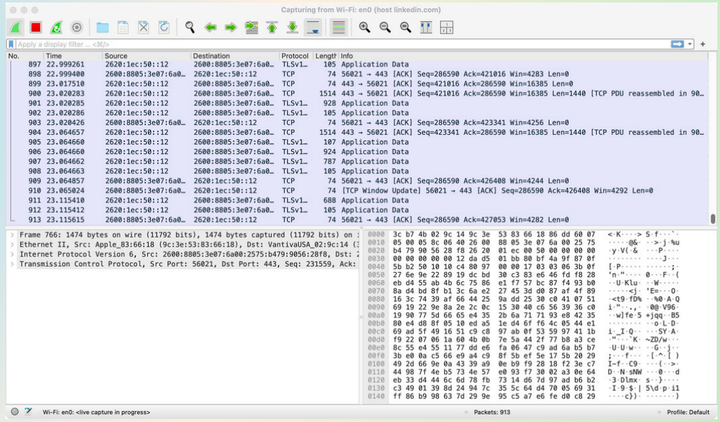
ChatGPT



Linkedin

Wireshark Collected Data





Exported Data

	Packet Count	Total Length	Avg Interval (s)	Max Interval (s)	Min Interval (s)	Avg Length (bytes)	Max Length (bytes)	Min Length (bytes)
2	16	2194	3.829345266666667	44.641063	9.200000000220145e-05	137.125	295	86
3	2603	1507206	0.0120919903920061	1.2703659999999992	0.0	579.026507875282	3660	54
4	8	958	0.6930795714285714	4.829743000000001	0.0002279999999998	119.75	193	74
5	2384	1226273	0.00959108015107	1.390694	0.0	514.3762583892617	15894	54
6	729	225431	0.0684117403846153	1.699636	0.0	309.2331961591221	10274	54
7	2222	1434237	0.0065306866276452	0.902013	0.0	645.471197119712	6328	54
8	3724	3425019	0.0133191576685468	1.2733440000000016	0.0	919.7150912996776	14454	54
9	2858	1347411	0.0161667028351417	1.9926239999999995	0.0	471.452414275713	8808	54
10	2551	1706454	0.0167094509803921	1.0135560000000012	0.0	668.9353194825559	15405	54
11	8534	8053355	0.0040767602250087	1.2613500000000002	0.0	943.678814155144	15031	54
12	1883	900887	0.0144601535600425	1.15137	0.0	478.431757833244 8	16460	54
13	1173	748607	0.0147205042662116	1.5354919999999996	0.0	638.1986359761296	7861	54
14	1632	663671	0.024266287553648	1.423649000000001	0.0	406.6611519607843	4975	54
15	1058	698412	0.0068863727530747	0.469006	0.0	660.124763705104	3660	54
16	749	394171	0.0061770561497326	0.4155759999999997	0.0	526.2630173564753	3660	54

	Packet Count	Total Length	Avg Interval (s)	Max Interval (s)	Min Interval (s)	Avg Length (bytes)	Max Length (bytes)	Min Length (bytes)	Most
2	2309	1683228	0.01729396	4.163646999999999	0	728.9857081	1514	54	1514
3	3270	1970394	0.010998199143468952	3.298861	0	602.5669724770643	1442	74	74
4	2076	1279110	0.017192924337349397	7.995936999999998	0	616.1416184971098	1442	74	74
5	3032	2057493	0.009012682	3.8215850000000025	0	678.5926781002638	1442	74	1442
6	3447	2249018	0.010853846581807951	5.446913000000002	0	652.4566289527125	1442	74	74
7	2271	1416532	0.028115036563876653	4.159818999999999	0	623.748128577191	1442	74	74
8	2882	1891438	0.015921671988892744	6.360627000000001	0	656.2935461	1442	74	74
9	3381	2105797	0.014537354733727812	5.442701	0	622.8325939071281	1442	74	74
10	2905	1736438	0.070345153	13.08749499999999	0	597.7411359724613	1442	74	74
11	3747	2450586	0.029735695	9.347669999999994	0	654.0128102481985	1442	74	74
12	2995	2176315	0.010454924181696728	5.495728999999999	0	726.6494156928213	1514	54	1514
13	3844	2591246	0.017516295	3.983114999999999	0	674.1014568158168	1514	54	54
14	4187	2861188	0.021852565934065933	5.687309999999965	0	683.3503701934559	1514	54	54
15	3023	2237499	0.018864486101919258	4.174109000000001	0	740.1584518690044	1514	54	1514
16	4574	3329542	0.015353316	4.421101	0	727.9278531	1514	54	1514

03 FEATURES EXTRACTION

Visited Website

Wireshark Collected Data

ChatGPT Trafic Data

No.	Time	Source	Destination	Protocol	Length	Info	Label
1	0	100.86.228.122	104.18.33.45	TCP	78	64744 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=3890593810 TSecr=0 SACK_PERM	ChatGPT
2	0.001022	100.86.228.122	104.18.33.45	TCP	78	64745 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=2360831628 TSecr=0 SACK_PERM	ChatGPT
3	0.001736	100.86.228.122	104.18.33.45	TCP	78	64746 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=2888318335 TSecr=0 SACK_PERM	ChatGPT
4	0.001813	100.86.228.122	104.18.33.45	TCP	78	64747 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=1053467278 TSecr=0 SACK_PERM	ChatGPT
5	0.003856	100.86.228.122	104.18.33.45	TCP	78	64748 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=2640338566 TSecr=0 SACK_PERM	ChatGPT
6	0.004103	100.86.228.122	104.18.33.45	TCP	78	64749 > 443 [SYN] Seq=0 Win=65535 Len=0 MSS=1460 WS=64 TSval=3737658740 TSecr=0 SACK_PERM	ChatGPT
7	0.017701	104.18.33.45	100.86.228.122	TCP	74	443 > 64747 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=3755523756 TSecr=1053467278 WS=8192	ChatGPT
8	0.017701	104.18.33.45	100.86.228.122	TCP	74	443 > 64744 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=4220765286 TSecr=3890593810 WS=8192	ChatGPT
9	0.017702	104.18.33.45	100.86.228.122	TCP	74	443 > 64745 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=1768673014 TSecr=2360831628 WS=8192	ChatGPT
10	0.017702	104.18.33.45	100.86.228.122	TCP	74	443 > 64746 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=1298033146 TSecr=2888318335 WS=8192	ChatGPT
11	0.017769	100.86.228.122	104.18.33.45	TCP	66	64747 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=1053467294 TSecr=3755523756	ChatGPT
12	0.017814	100.86.228.122	104.18.33.45	TCP	66	64744 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=3890593827 TSecr=4220765286	ChatGPT
13	0.017832	100.86.228.122	104.18.33.45	TCP	66	64745 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=2360831644 TSecr=1768673014	ChatGPT
14	0.017847	100.86.228.122	104.18.33.45	TCP	66	64746 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=2888318351 TSecr=1298033146	ChatGPT
15	0.017953	100.86.228.122	104.18.33.45	TCP	1454	64747 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=1053467294 TSecr=3755523756 [TCP PDU reassembled in 16]	ChatGPT
16	0.017966	100.86.228.122	104.18.33.45	TLSv1.3	494	Client Hello (SNI=openai.com)	ChatGPT
17	0.018117	100.86.228.122	104.18.33.45	TCP	1454	64744 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=3890593827 TSecr=4220765286 [TCP PDU reassembled in 18]	ChatGPT
18	0.018123	100.86.228.122	104.18.33.45	TLSv1.3	398	Client Hello (SNI=openai.com)	ChatGPT
19	0.018251	100.86.228.122	104.18.33.45	TCP	1454	64745 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=2360831644 TSecr=1768673014 [TCP PDU reassembled in 20]	ChatGPT
20	0.018257	100.86.228.122	104.18.33.45	TLSv1.3	462	Client Hello (SNI=openai.com)	ChatGPT
21	0.018385	100.86.228.122	104.18.33.45	TCP	1454	64746 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=2888318351 TSecr=1298033146 [TCP PDU reassembled in 22]	ChatGPT
22	0.01839	100.86.228.122	104.18.33.45	TLSv1.3	494	Client Hello (SNI=openai.com)	ChatGPT
23	0.01984	104.18.33.45	100.86.228.122	TCP	74	443 > 64749 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=1361514557 TSecr=3737658740 WS=8192	ChatGPT
24	0.01984	104.18.33.45	100.86.228.122	TCP	74	443 > 64748 [SYN, ACK] Seq=0 Ack=1 Win=65535 Len=0 MSS=1400 SACK_PERM TSval=222912174 TSecr=2640338566 WS=8192	ChatGPT
25	0.019878	100.86.228.122	104.18.33.45	TCP	66	64749 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=3737658757 TSecr=1361514557	ChatGPT
26	0.019902	100.86.228.122	104.18.33.45	TCP	66	64748 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=0 TSval=2640338583 TSecr=222912174	ChatGPT
27	0.019992	100.86.228.122	104.18.33.45	TCP	1454	64749 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=3737658757 TSecr=1361514557 [TCP PDU reassembled in 28]	ChatGPT
28	0.020001	100.86.228.122	104.18.33.45	TLSv1.3	494	Client Hello (SNI=openai.com)	ChatGPT
29	0.020115	100.86.228.122	104.18.33.45	TCP	1454	64748 > 443 [ACK] Seq=1 Ack=1 Win=131840 Len=1388 TSval=2640338583 TSecr=222912174 [TCP PDU reassembled in 30]	ChatGPT
30	0.020119	100.86.228.122	104.18.33.45	TLSv1.3	462	Client Hello (SNI=openai.com)	ChatGPT

Extracted Features

- Packet Count
- Total Length
- Avg Packet Interval
- Max Packet Interval
- Min Packet Interval
- Avg Packet Length
- Max Packet Length
- Min Packet Length
- Most Common Length

03 FEATURES EXTRACTION

Dataset Size

Label	Size
ChatGPT	20
LinkedIn	20
Wikipedia	30
Reddit	15
TOTAL	**85**

Dataset Samples

Avg Interval (s)	Max Interval (s)	Min Interval (s)	Avg Length (bytes)	Max Length (bytes)	Min Length (bytes)	Most Common Length (bytes)	Label
0.0382649160899654	4.988439999999997	0.0	782.9887640449438	1434	86	1434	ChatGPT
0.03762468567251462	9.353463000000001	0.0	737.1008035062089	1434	86	1434	ChatGPT
0.07055149416342413	15.034570000000002	0.0	699.6101118133204	1434	86	1434	ChatGPT
0.07024142705167173	5.140540000000001	0.0	639.5720789074355	1434	86	1434	ChatGPT
0.0844454889310563	5.0997270000000015	0.0	535.9329962073325	1434	86	86	ChatGPT
0.06552765693430657	2.516216	0.0	615.4805825242719	1434	86	1434	ChatGPT
0.023048338441039307	3.2136790000000026	0.0	770.5912117177097	1434	86	1434	ChatGPT
0.04525340731399748	6.5650209999999944	0.0	685.4631379962193	1434	86	1434	ChatGPT
0.017209531044926805	1.4559240000000004	0.0	491.3466195761857	1434	86	86	ChatGPT
0.08304730136986302	11.884944999999998	0.0	594.255859375	1434	86	86	ChatGPT
0.019915201943844493	4.6144940000000005	0.0	822.0215749730313	1434	86	1434	ChatGPT
0.012855877828054299	2.8242019999999997	0.0	418.9861464517953	1434	86	86	ChatGPT
0.033415478694469626	10.239505000000001	0.0	777.5126811594203	1434	86	1434	ChatGPT

Total Length	Avg Interval (s)	Max Interval (s)	Min Interval (s)	Avg Length (bytes)	Max Length (bytes)	Min Length (bytes)	Most Common Length (bytes)	Label
803559	0.046798	9.741225	0	994.503713	1514	74	1514	Wikipedia
1056001	0.482537	99.37193	0	855.754457	1514	74	1514	Wikipedia
352540	0.289238	82.263481	0	1061.86747	1514	74	1514	Wikipedia
638214	1.07554	136.02592	0	622.647805	1514	74	74	Wikipedia
1736902	0.185358	35.491703	0	798.942962	1514	54	1514	Wikipedia
1134192	0.688325	85.786671	0	634.33557	1514	66	1514	Wikipedia
67713	0.982695	45.015105	0	593.973684	1514	74	74	Wikipedia
592867	0.592949	45.006885	0	754.283715	1514	74	1514	Wikipedia
411863	0.330857	45.003131	0	1027.089776	1514	74	1514	Wikipedia
908268	0.532771	242.67024	0	805.202128	1514	74	1514	Wikipedia
1823145	0.225795	33.38485	0	800.678524	1514	74	1514	Wikipedia

We Split our data to Train, Test, and Validation in 6:2:2

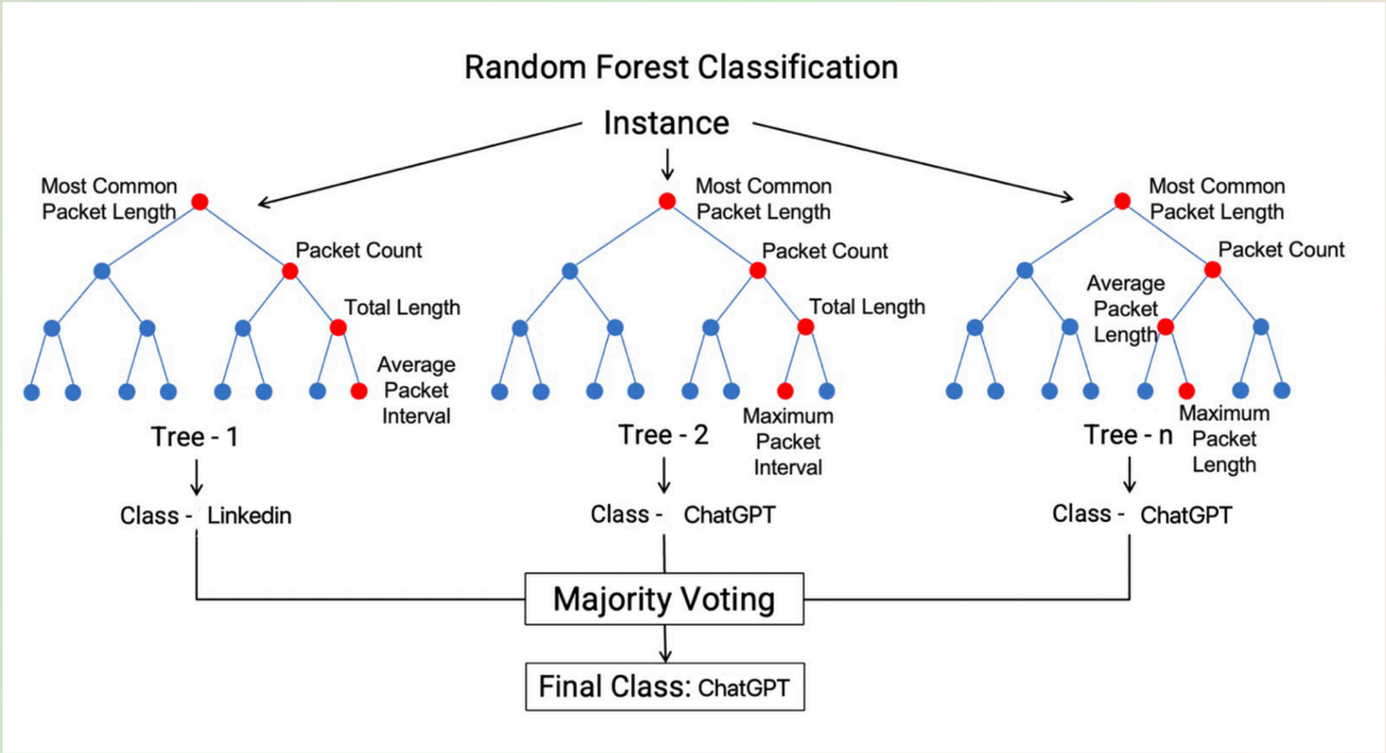
04. METHOD

RANDOM FORESTS

Definition: Random Forest is a supervised machine learning model that combines the results of **many decision trees**. Instead of relying on just one tree to make a prediction, it builds multiple trees and then **takes a vote from all of them.**

How it works:

1. Each data point includes features like packet count, total length, and packet intervals.
2. The model learns patterns from this data to predict which website—like ChatGPT, Reddit, Wikipedia, or LinkedIn—the traffic belongs to.



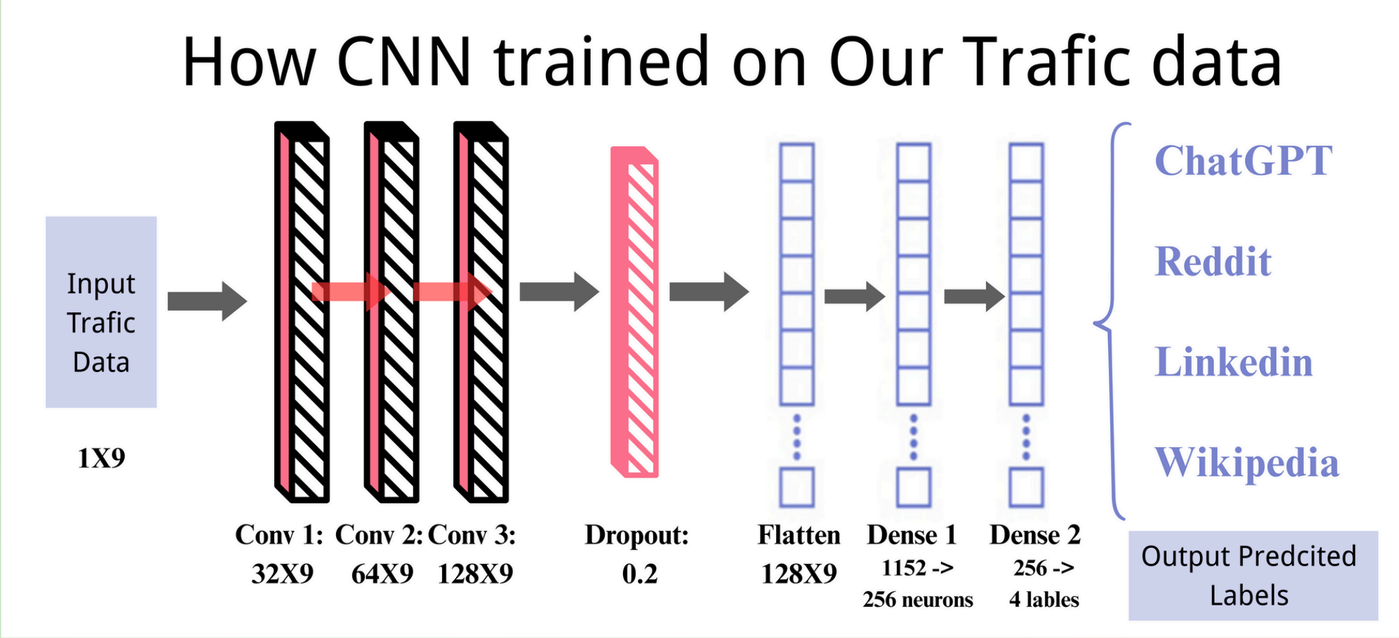
04. METHOD

CNN

Definition: CNN stands for Convolutional Neural Network. It's a type of deep learning model that is especially good at understanding patterns in data, like images or sequences. It uses **multiple layers of filters** to learn important features automatically. Here, we used **1D CNN model.**

How it works:

1. Each sample was a **sequence of network features** like packet count, packet length, and intervals.
2. The CNN processed this data through 3 convolution layers, 1 flatten, and 2 dense process and made a prediction about which website the traffic came from — ChatGPT, Reddit, Wikipedia, or LinkedIn.



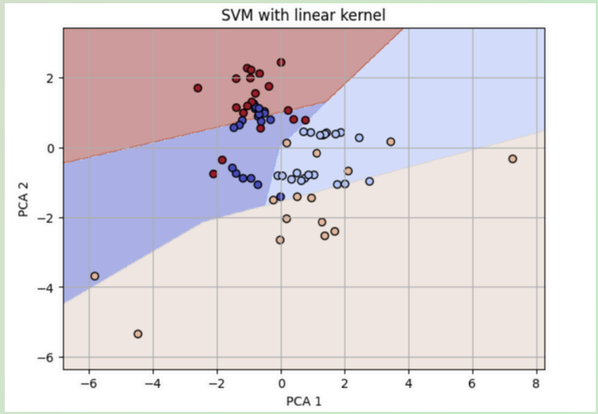
04. METHOD

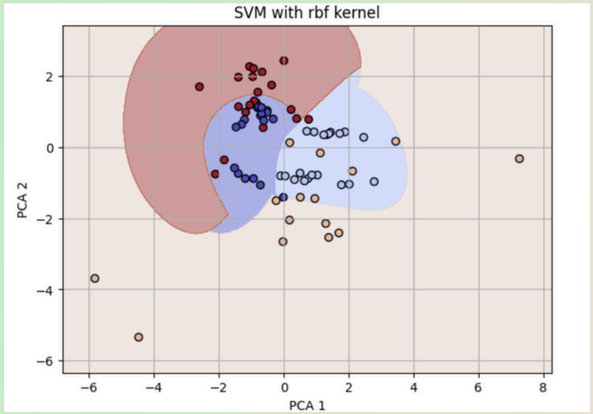
SVM

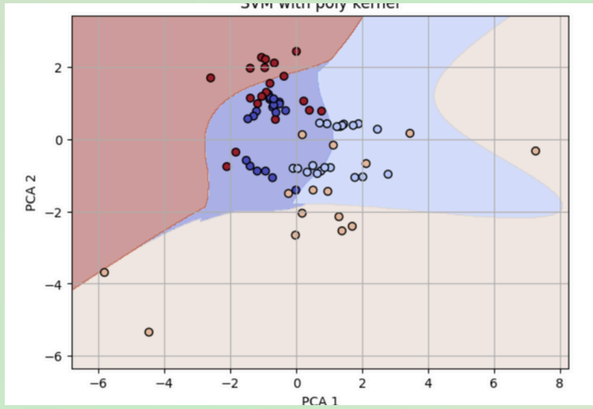
Definition: A Support Vector Machine (SVM) uses Support Vector Classifier (SVC) to draw decision boundaries that separate data points into classes. It finds optimal weights and bias terms, forming a linear (or non-linear) decision boundary based on feature vectors. SVC can apply different kernels to define the shape of this boundary depending on the complexity of the data.

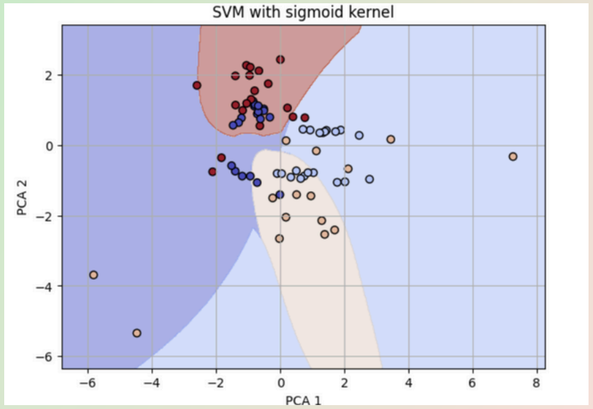
How it works:

1. **SVC uses parameters like C to control the trade-off between maximizing the margin and minimizing classification errors.**
2. **Kernels: Linear, Poly (Curved boundaries using polynomials), RBF (Radial Basis Function) uses flexible boundary for non-linear separation, Sigmoid (good for binary classification)**
3. **A GridSearch is often used to tune hyperparameters (like C, gamma, and kernel type) via cross-validation.**









04. METHOD

LINEAR REGRESSION

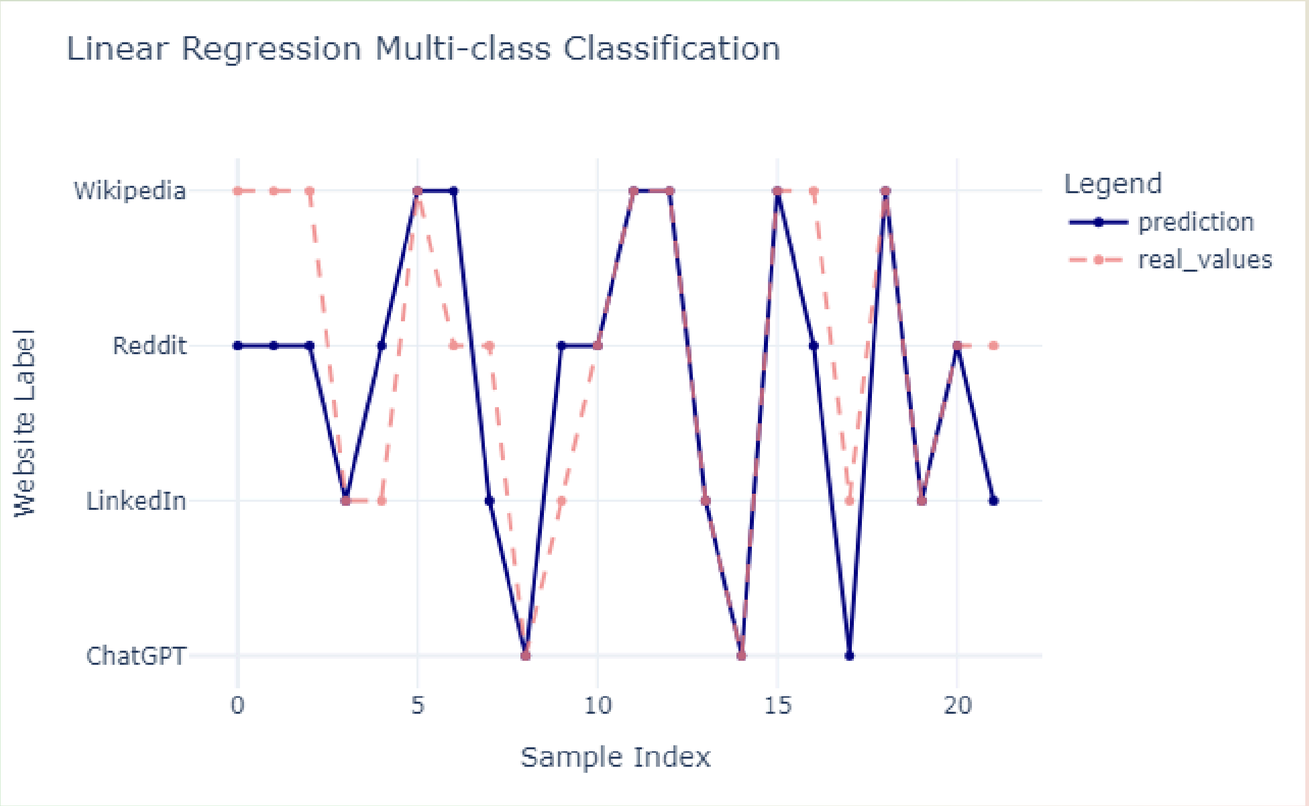
Definition: Utilized as a basic baseline model for comparison, acknowledging its inherent limitations for classification tasks.

How it works:

Generated continuous predictions which were then rounded to the nearest integer and clipped to the valid label range (0-3) to assign discrete class membership for the categories.

Baselines:

- Achieved an approximate test accuracy of 55%.
- Reached a Macro F1 score of 0.58.
- Showed significant performance disparities across classes, indicating poor discriminatory power, particularly for Reddit (F1: 0.31) and LinkedIn (F1: 0.55).



04. METHOD

LOGISTIC REGRESSION

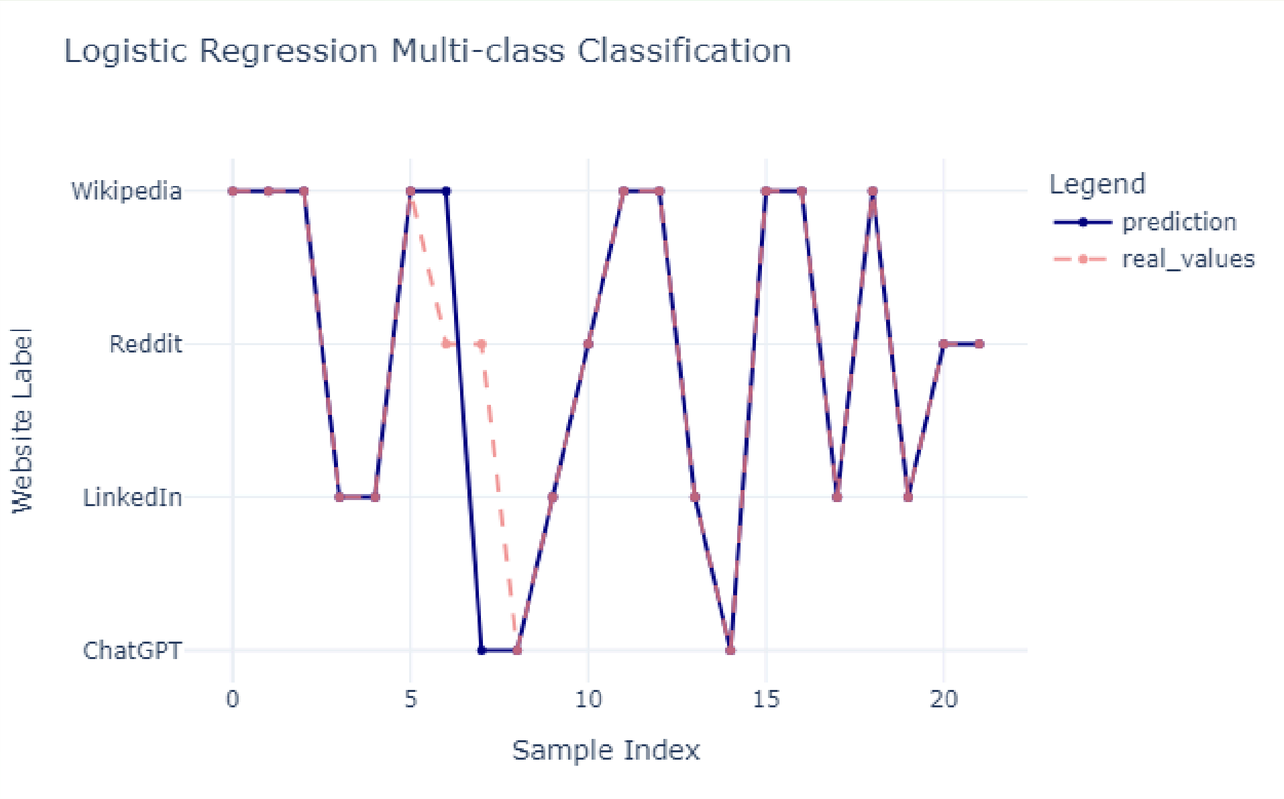
Definition: Implemented as a standard classification model, expected to be more suitable for distinguishing between the four categories.

How it works:

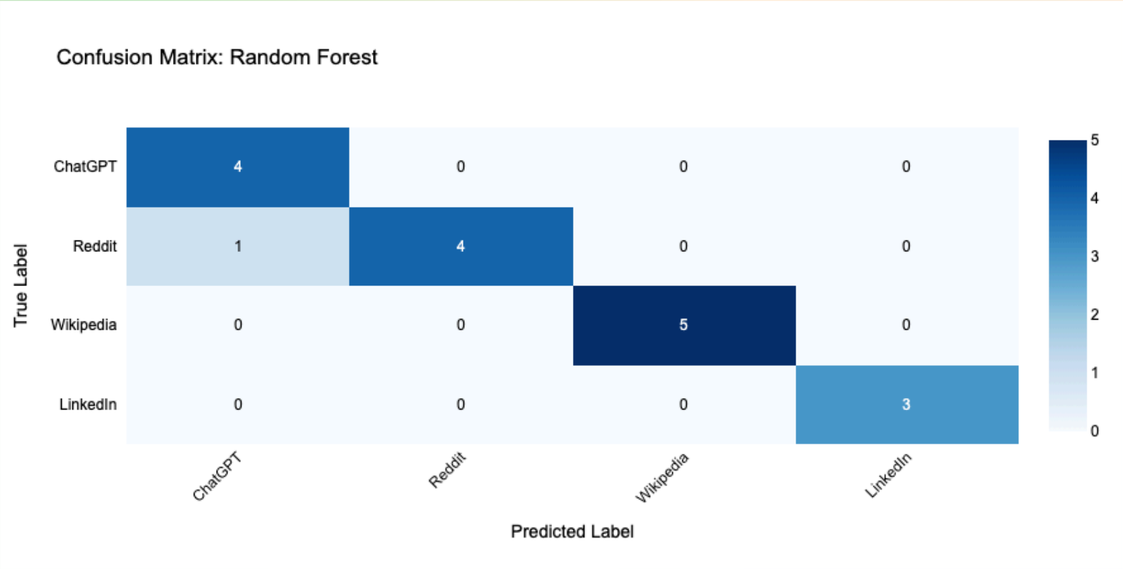
Applied standard algorithm principles to classify data based on learned feature relationships after data preparation (handling missing values, encoding labels).

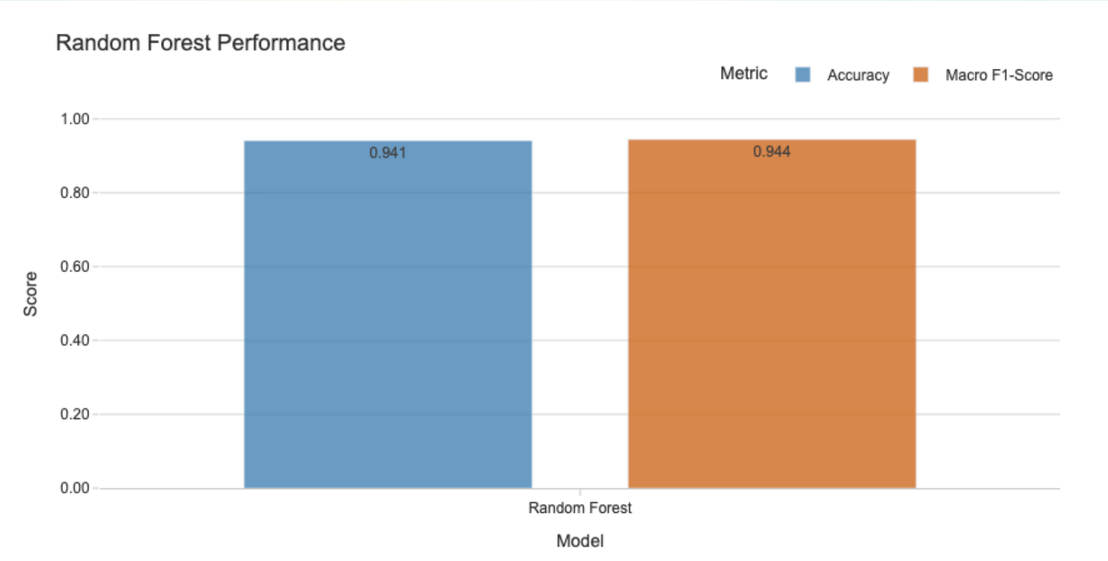
Baselines:

- Achieved a high test accuracy of 91%.
- Performed excellently for LinkedIn (F1: 1.00) and Wikipedia (F1: 0.95).
- Showed perfect recall for ChatGPT (F1: 0.80).
- Exhibited lower recall for the Reddit class (0.60), although overall precision and recall were high across most classes.



05 RESULTS *RANDOM FORESTS*

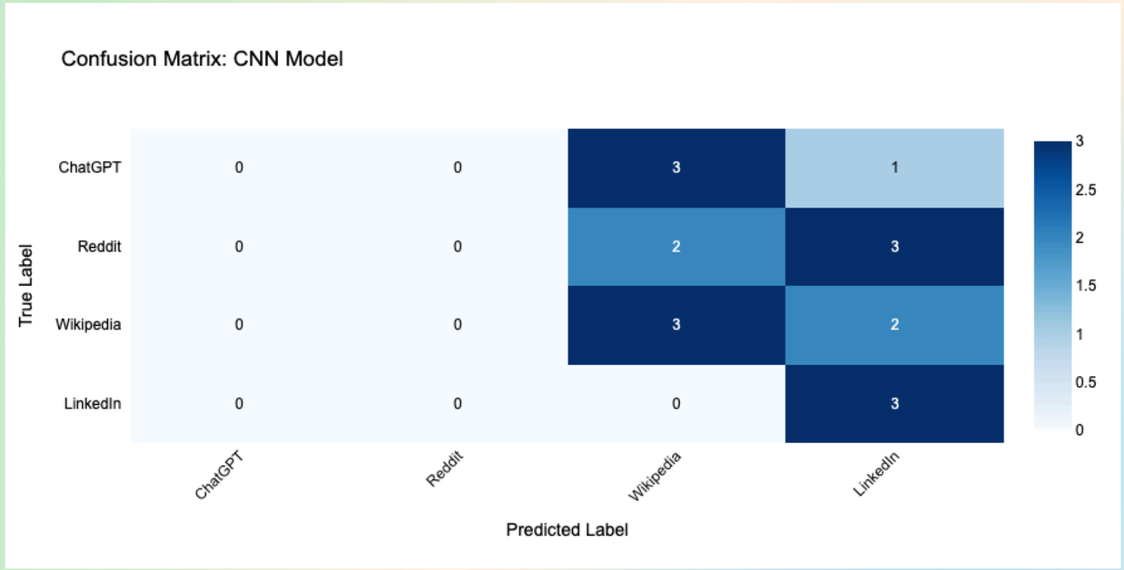


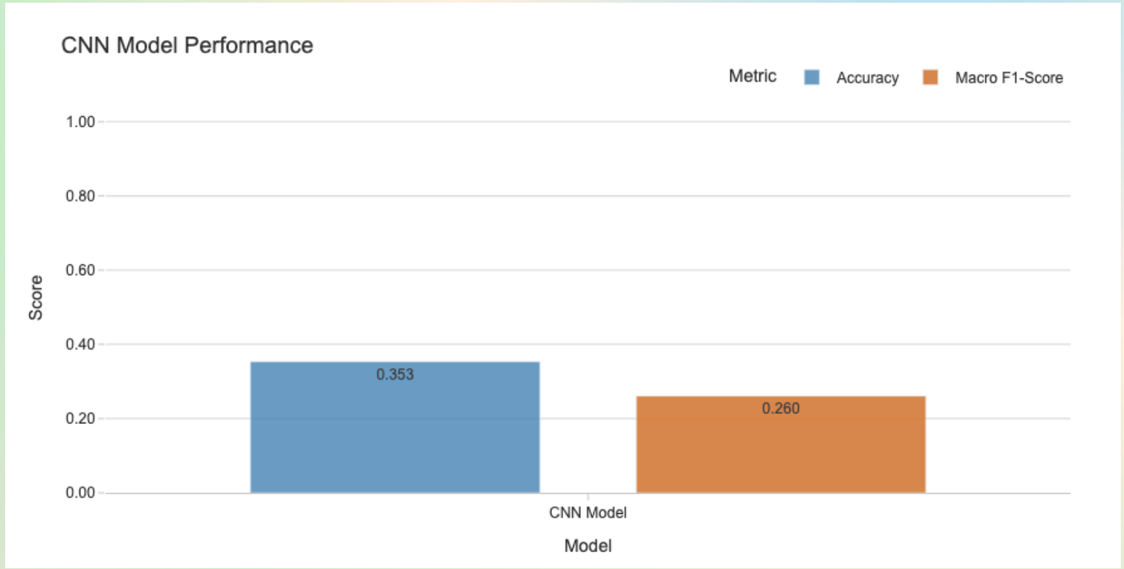


Accuracy in Test data: 94.1%

Macro F1 in Test data: 94.4%

05 RESULTS *CNN*

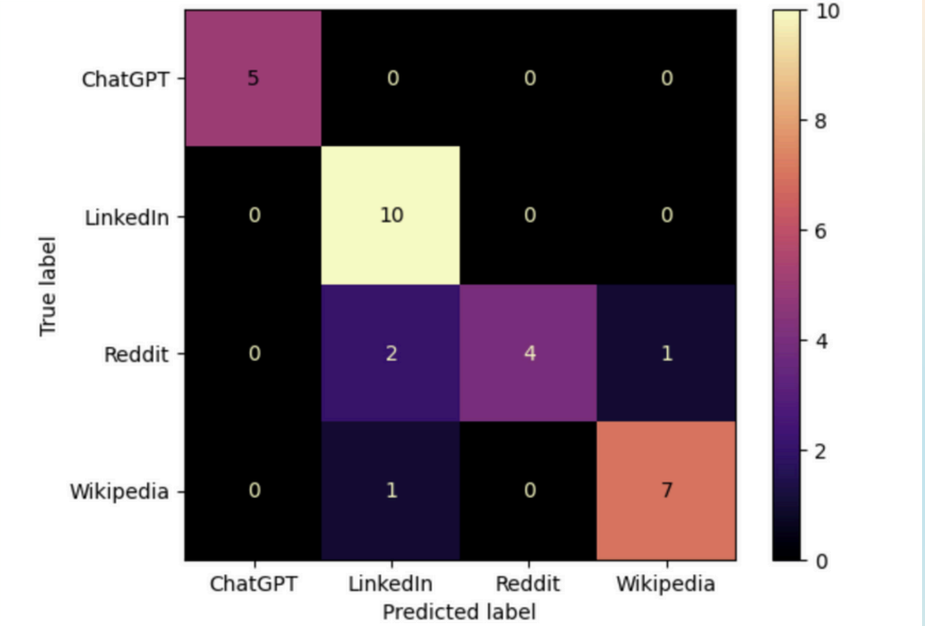


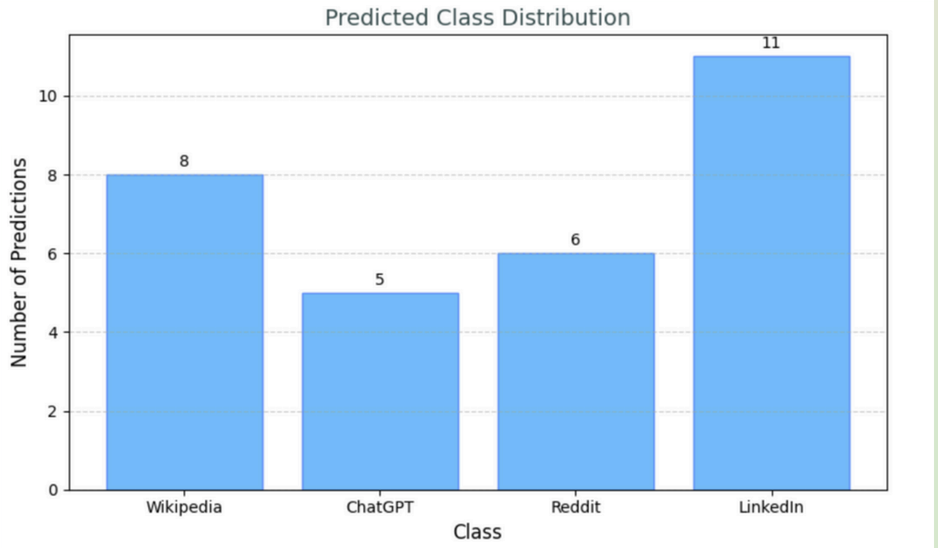


Accuracy: 35.3%

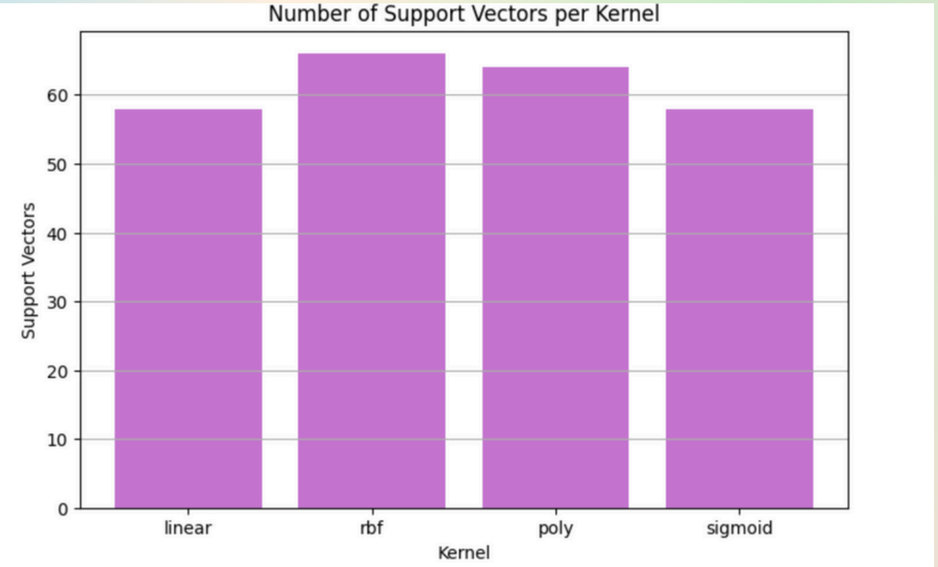
Macro F1: 26%

05 RESULTS *SVM*






```
Best parameters: {'C': 1, 'degree': 1, 'gamma': 'scale', 'kernel': 'rbf'}
Best CV score: 0.9111

Test set classification report:
              precision    recall  f1-score   support

     ChatGPT       1.00      1.00      1.00         5
    LinkedIn       0.77      1.00      0.87        10
      Reddit       1.00      0.57      0.73         7
   Wikipedia       0.88      0.88      0.88         8

    accuracy                           0.87        30
   macro avg       0.91      0.86      0.87        30
weighted avg       0.89      0.87      0.86        30
```



05 RESULTS *SVM*

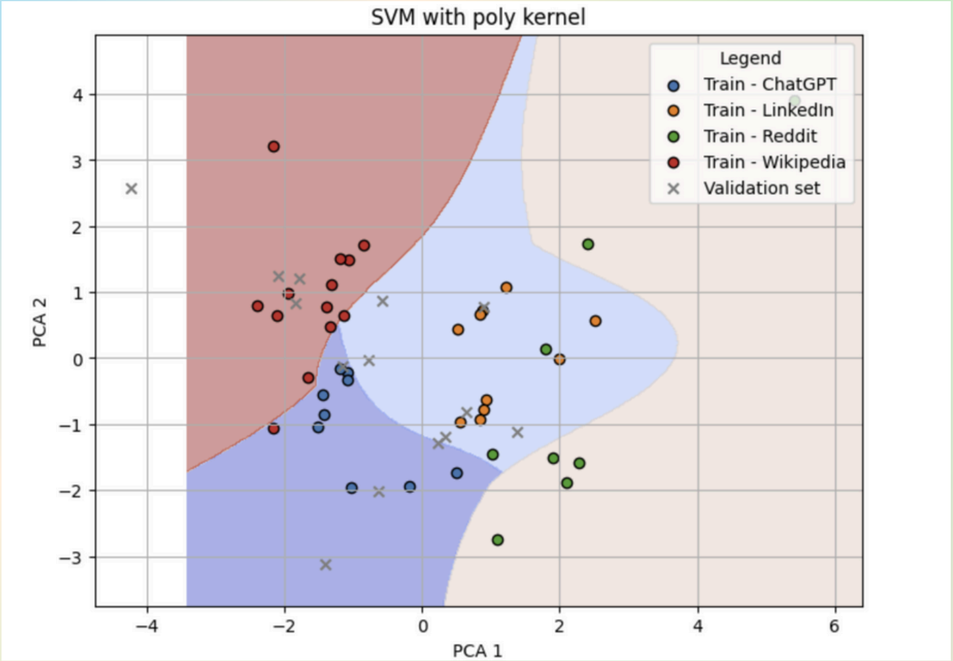
POLY Kernel: Degree=1, C=10, Test Accuracy: 0.9333

SIGMOID Kernel: Degree=1, C=1, Test Accuracy: 0.9000

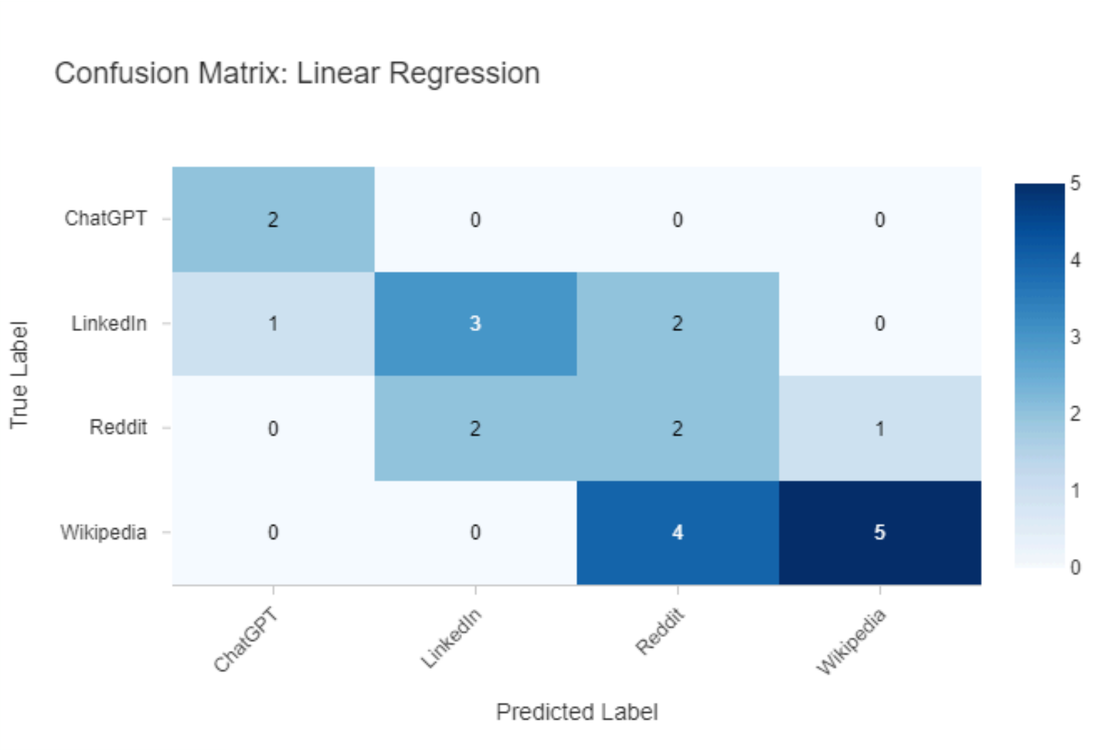
LINEAR Kernel: Degree=1, C=1, Test Accuracy: 0.9333

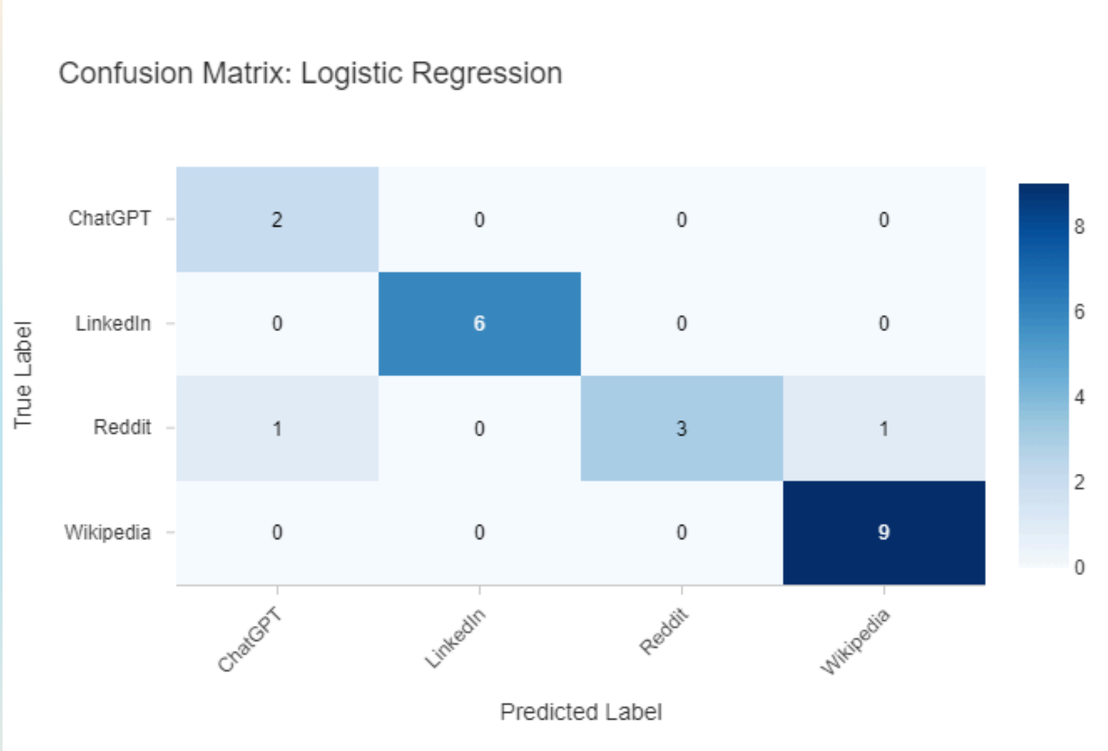
Gamma + RBF Kernel: Best gamma: 0.0027000000000000001, Best C: 100, Best Accuracy: 0.9667

```
Fitting 5 folds for each of 240 candidates, totalling 1200 fits
Best parameters: {'C': 10, 'degree': 1, 'gamma': 'scale', 'kernel': 'poly'}
Best CV score (on Train): 1.0000

Validation set accuracy: 0.8571
              precision    recall  f1-score   support

     ChatGPT       0.75      1.00      0.86         3
    LinkedIn       0.80      1.00      0.89         4
      Reddit       0.00      0.00      0.00         2
   Wikipedia       1.00      1.00      1.00         5

    accuracy                           0.86        14
   macro avg       0.64      0.75      0.69        14
weighted avg       0.75      0.86      0.79        14


Test set accuracy: 0.9286
              precision    recall  f1-score   support

     ChatGPT       1.00      1.00      1.00         3
    LinkedIn       1.00      1.00      1.00         4
      Reddit       1.00      0.67      0.80         3
   Wikipedia       0.80      1.00      0.89         4

    accuracy                           0.93        14
   macro avg       0.95      0.92      0.92        14
weighted avg       0.94      0.93      0.93        14
```

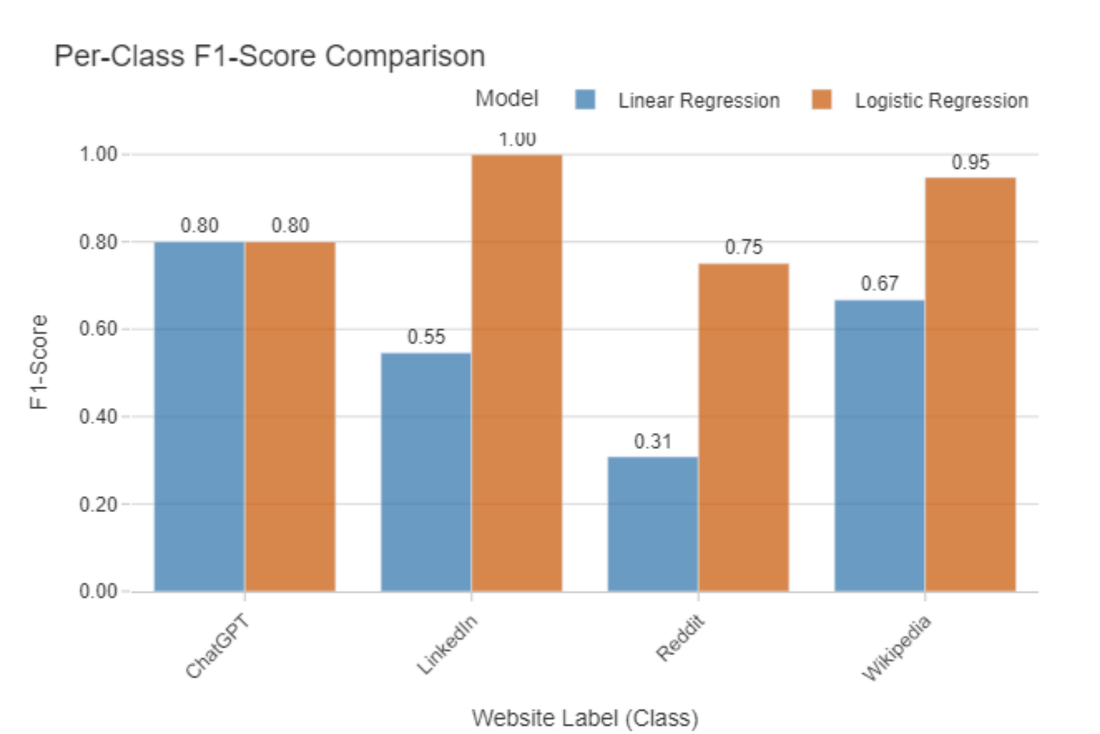


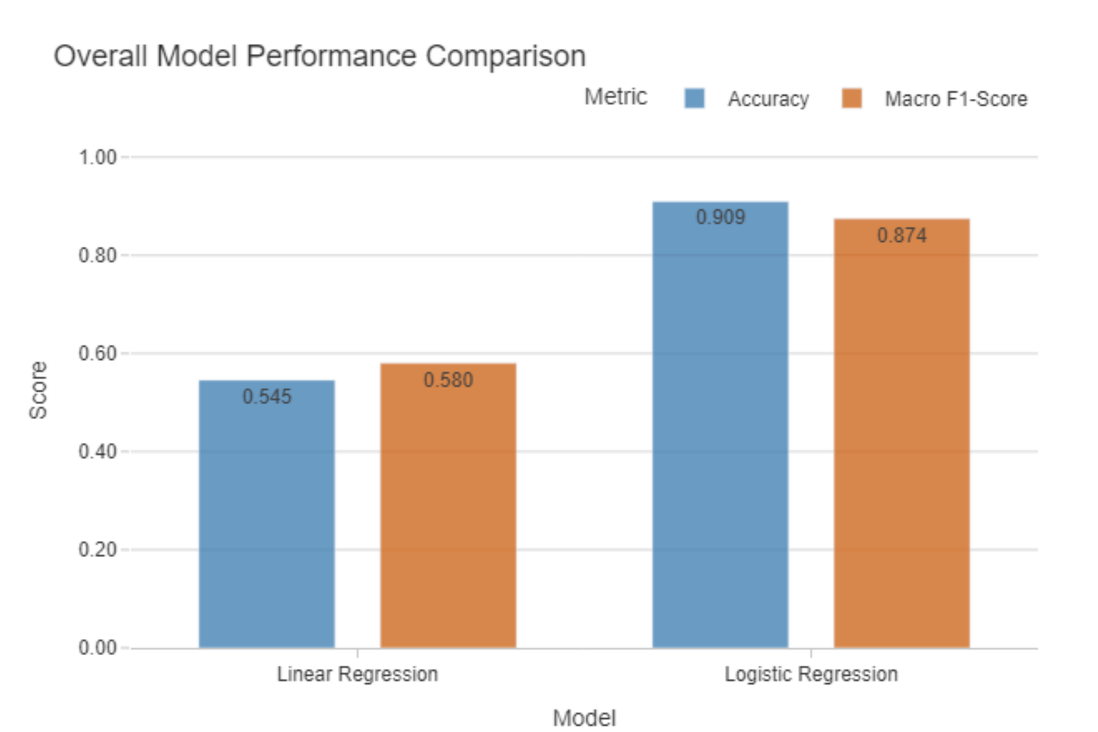
05 RESULTS *LINEAR AND LOGISTIC REGRESSION*





05 RESULTS *LINEAR AND LOGISTIC REGRESSION*





Comparison of Model Performance

Model	Accuracy	Macro F1 Score
Linear Regression	0.54	0.58
Logistic Regression	0.91	0.87
CNN	0.35	0.26
SVM	**0.97**	**0.87**
Random Forest	0.94	0.94

06 | FUTURE DISCUSSIONS

- **Datasets Size**
- **Model Complexity**
- **Real-Time Network Trafic**

THANK YOU

Hosanna Root, Evan Abney, Jinyan Kuang, Abhayprad Jha, Krishika Pudasaini