ASSIGNMENT 3 - INLP

REPORT

SKIP GRAM WITH NEGATIVE SAMPLING

ANALYSIS WHEN WINDOW SIZE = 10

ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =10

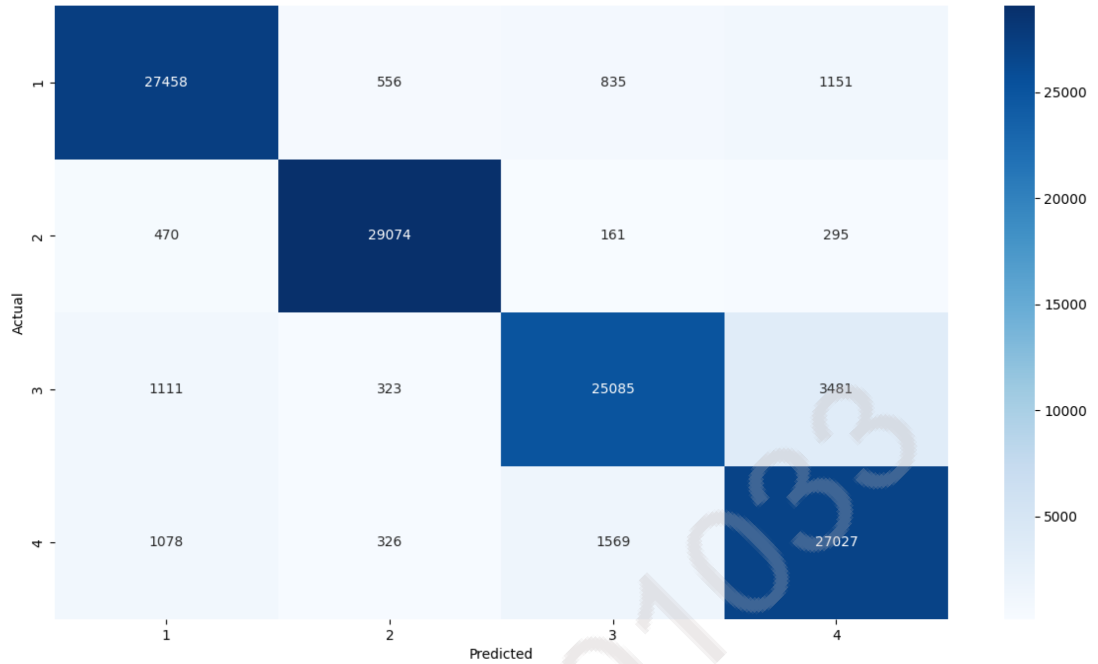
```
Test Accuracy: 88.82%
Training Set Performance:
Accuracy: 0.91
F1 Score: 0.91
Precision: 0.91
Recall: 0.91
Confusion Matrix:

Test Set Performance:
Accuracy: 0.89
F1 Score: 0.89
Precision: 0.89
Recall: 0.89
Confusion Matrix:
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERABAD\sem
```


- The calculated metrics substantiate the model's efficiency, evidenced by high accuracy scores of 0.91 for the training set and 0.89 for the testing set.
- The F1 score, an important metric that balances precision and recall, stands at 0.91 for training and 0.89 for testing.
- The precision metric, which reflects the model's ability to label as positive only those samples that are actually positive, aligns with the F1 score at 0.91 for training and 0.89

for testing.

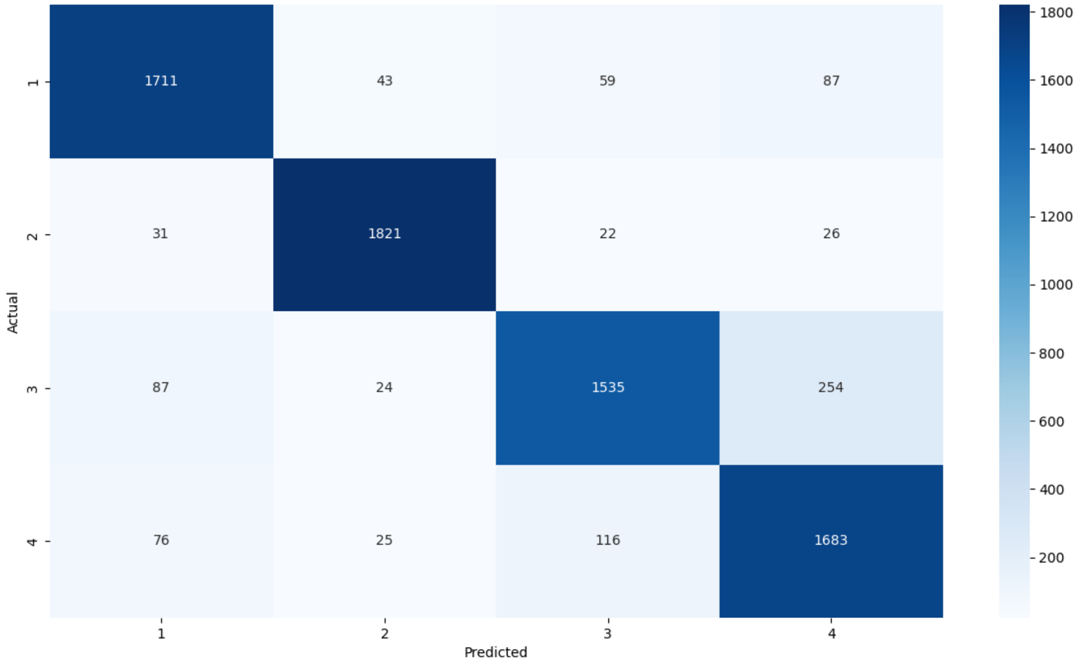
- The consistency of these metrics across training and testing sets indicates that model has performed well. **But then if you closely observe the test set accuracy, it is less compared to what we achieved in window size = 5. which indicates the possibility of noise being captured as context. However, for 10 we dint observe much, but for sure if we increase further, it will definitely decrease the accuracy further.**

<u>CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =10</u>



- The training data shows a comparable trend with strong true positive rates (27458 for Class 1, 29074 for Class 2, 25085 for Class 3, and 27027 for Class 4).
- Misclassifications are present but relatively low, indicating the model's aptitude for learning from the training data.

<u>CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =10</u>



- The diagonal numbers, representing correct predictions, are high for all classes (1711 for Class 1, 1821 for Class 2, 1535 for Class 3, and 1683 for Class 4).
- The misclassification rates have marginally increased in comparison to the window size of 5, suggesting a slightly reduced generalization capability.

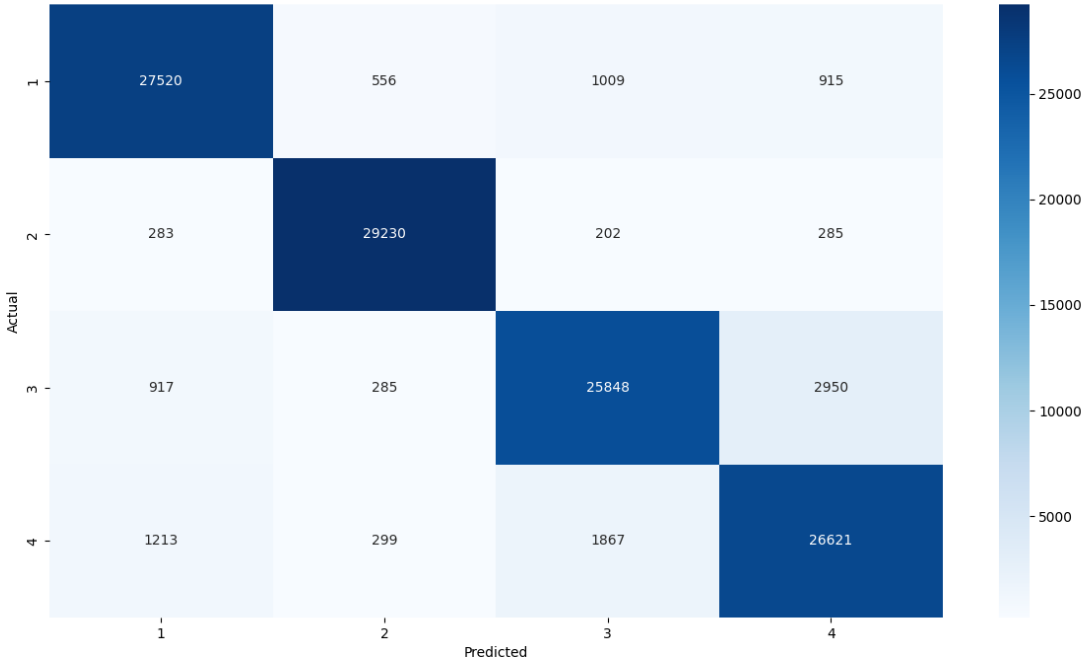
ANALYSIS WHEN WINDOW SIZE = 5

ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =5

```
Test Accuracy: 89.55%
Training Set Performance:
Accuracy: 0.91
F1 Score: 0.91
Precision: 0.91
Recall: 0.91
Confusion Matrix:

Test Set Performance:
Accuracy: 0.90
F1 Score: 0.90
Precision: 0.90
Recall: 0.90
Confusion Matrix:
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERABAD\sem2\NLP\20232010
```

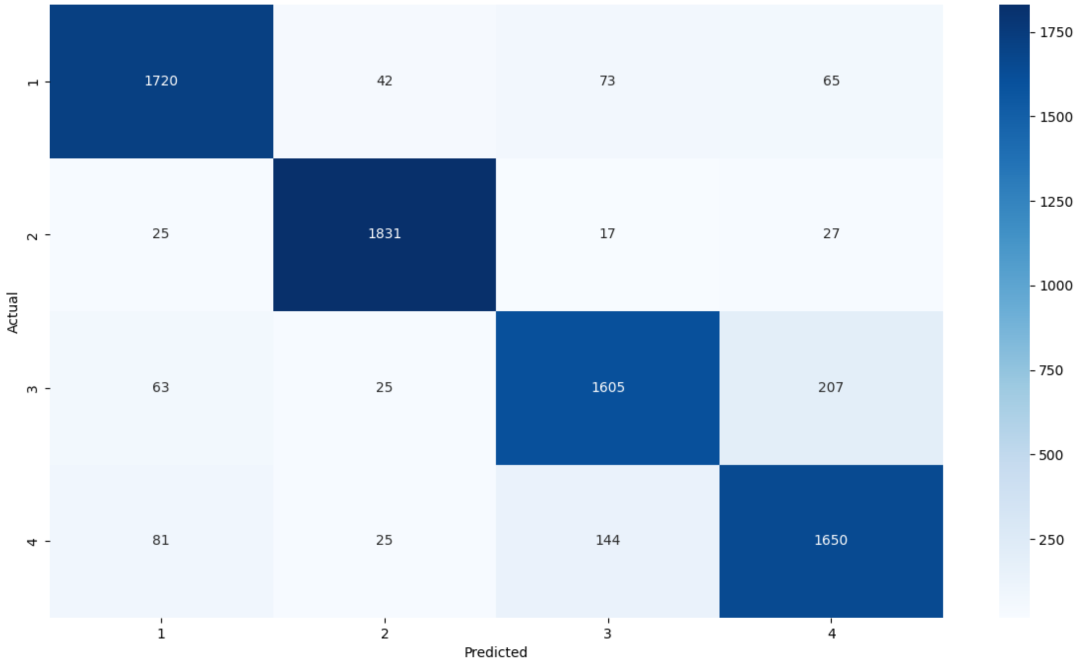
- The overall performance metrics with an accuracy of 0.91 and an F1 score of 0.91 for the training set, alongside an accuracy of 0.90 and an F1 score of 0.90 for the testing set.
- These figures imply a balanced performance in the model's prediction capabilities.
- The precision and recall metrics for both training and testing are in at 0.91 and 0.90, respectively, showing that the model is actually able to identify the positive instances properly.
- **If you closely observe, the best accuracy for test set which is around 90 percent, we have got when context window is 5. suggests, that window size should not be that long or very small.**

<u>**CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =5**</u>



- The training data shows a comparable trend with strong true positive rates (27520 for Class 1, 29230 for Class 2, 25848 for Class 3, and 26621 for Class 4).
- Misclassifications are present but relatively low, indicating the model's aptitude for learning from the training data.

CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =5



- For the testing data, the confusion matrix shows a strong predictive accuracy as stated above.
- The diagonal numbers, representing correct predictions, are high for all classes (1720 for Class 1, 1831 for Class 2, 1605 for Class 3, and 1650 for Class 4).

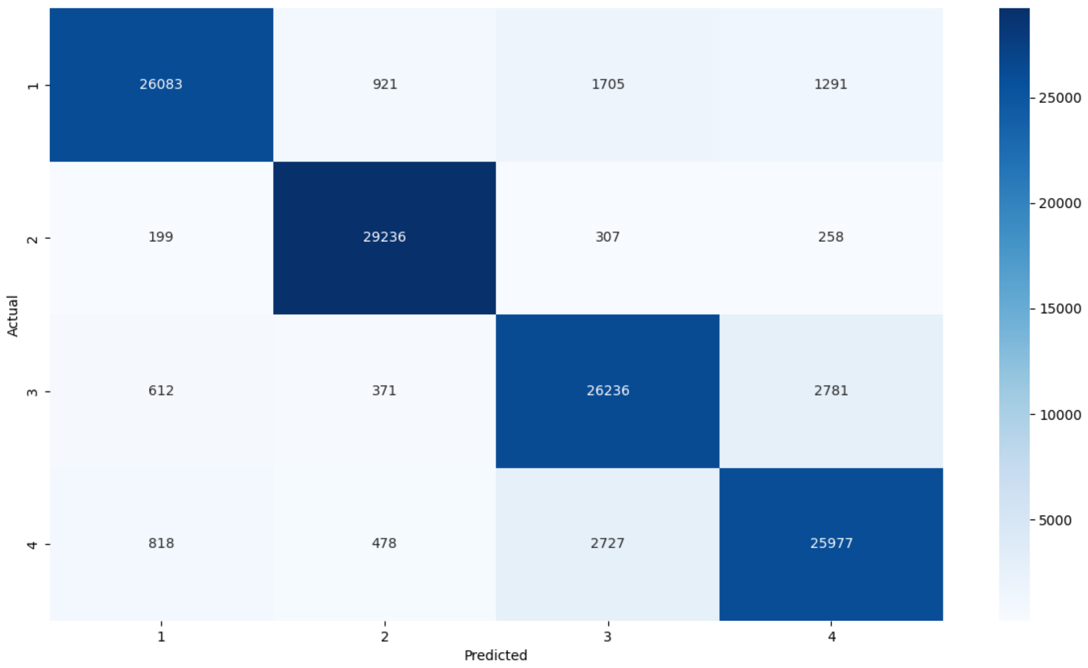
<u>ANALYSIS WHEN WINDOW SIZE = 2</u>

<u>ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =2</u>

```
Test Accuracy: 87.68%
Training Set Performance:
Accuracy: 0.90
F1 Score: 0.90
Precision: 0.90
Recall: 0.90
Confusion Matrix:

Test Set Performance:
Accuracy: 0.88
F1 Score: 0.88
Precision: 0.88
Recall: 0.88
Confusion Matrix:
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERABAD\sem2\NLP\202320
```

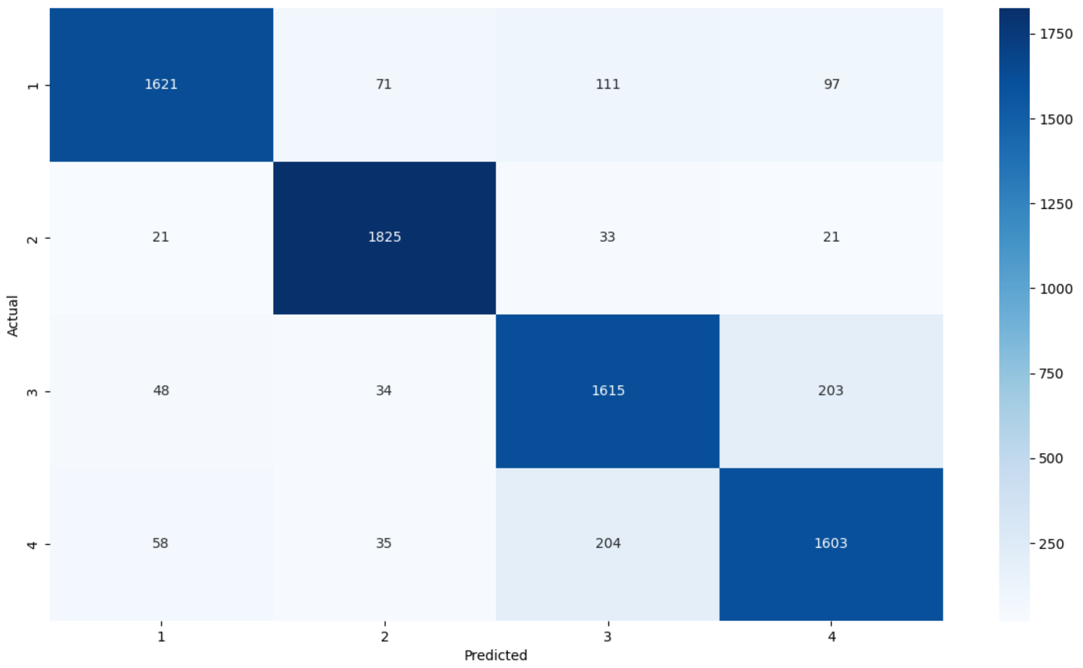
- The analysis for a window size of 2 points to a capable model with a good understanding of the task at hand.
- However, there is a clear indication that the limited contextual window impacts the model's performance, particularly its ability to differentiate between closely related classes.
- Overall Performance metrics indicate a near-consistent level of precision and recall across training and testing sets, with training scores at 0.90 for accuracy, F1 score, precision, and recall, and testing scores at 0.88 for the same metrics.

<u>CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =2</u>



- The training set reveals a strong capacity for generalization, with true positives at 26083, 29236, 26236, and 25977 for the respective classes.
- However, as with the testing set, the rate of misclassification is higher than what was observed with larger windows, implying **that more contextual information could enhance the model's predictive accuracy**.

CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =2



- The confusion matrix for the testing data with a window size of 2 presents high classification accuracy, with true positives recorded at 1621, 1825, 1615, and 1603 for classes 1 to 4, respectively.
- Despite the high accuracy, there's a noticeable amount of misclassification, suggesting that the **reduced context window may limit the model's understanding of broader linguistic structures.**

OVERALL CONCLUSION ABOUT SKIP GRAM METHOD

- Through comparative analysis across varying window sizes, it becomes evident that the window size has a tangible impact on the model's performance.
- Larger window sizes generally lead to higher accuracy and F1 scores, indicating that additional context enhances the model's predictive capabilities. **At the same time it also tells that a window size after a threshold can capture noise as the context and hence performance may deteriorate in the case of Skip gram.**
- However, even with smaller window size, the model demonstrates respectable generalization, as evidenced by its consistent precision and recall across training and testing datasets.
- The slight performance decline when using smaller window sizes highlights the importance of a sufficiently broad context for the model to capture essential linguistic patterns.

- In conclusion, while the model is robust and performs effectively across different configurations, **optimizing the context window size is crucial.**

SINGULAR VALUE DECOMPOSITION

ANALYSIS WHEN WINDOW SIZE = 10

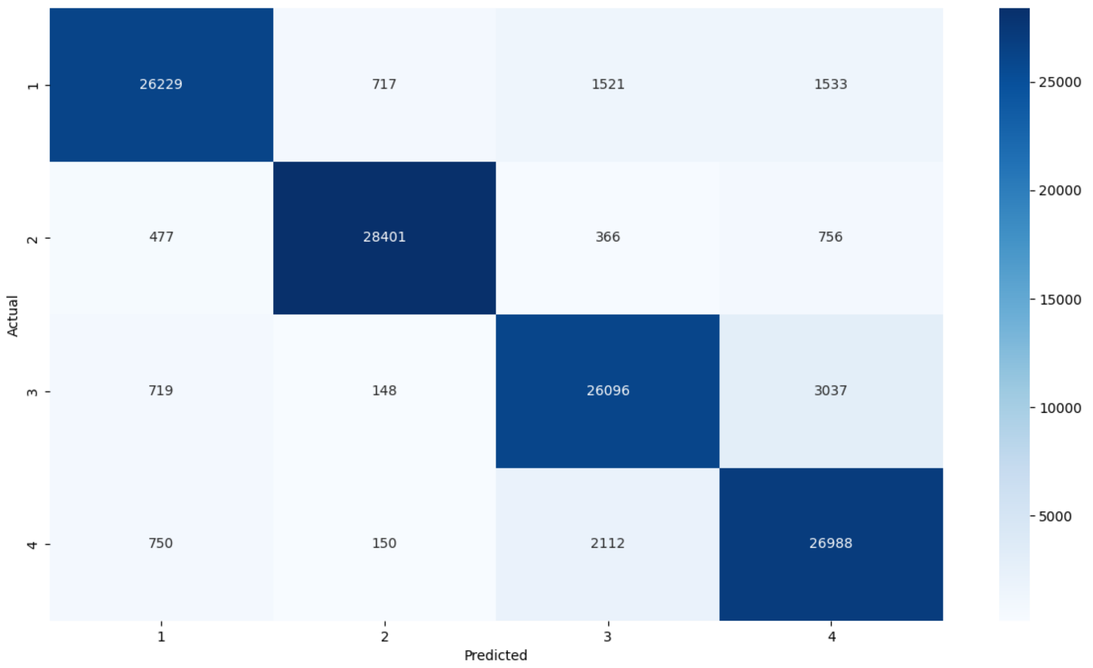
ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =10

```
Test Accuracy: 87.05%
Training Set Performance:
Accuracy: 0.90
F1 Score: 0.90
Precision: 0.90
Recall: 0.90

Test Set Performance:
Accuracy: 0.87
F1 Score: 0.87
Precision: 0.87
Recall: 0.87
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERABAD\sem2\NLP\2023201033_A3>
```


- The overall metrics display accuracy, F1 score, precision, and recall and is exhibiting high values (0.87), which proves that the model has learned to generalize well from the training data to unseen data.
- Since, the testing metrics are on par with the training ones, it is demonstrating the model's excellent generalization capabilities even with the extended context.

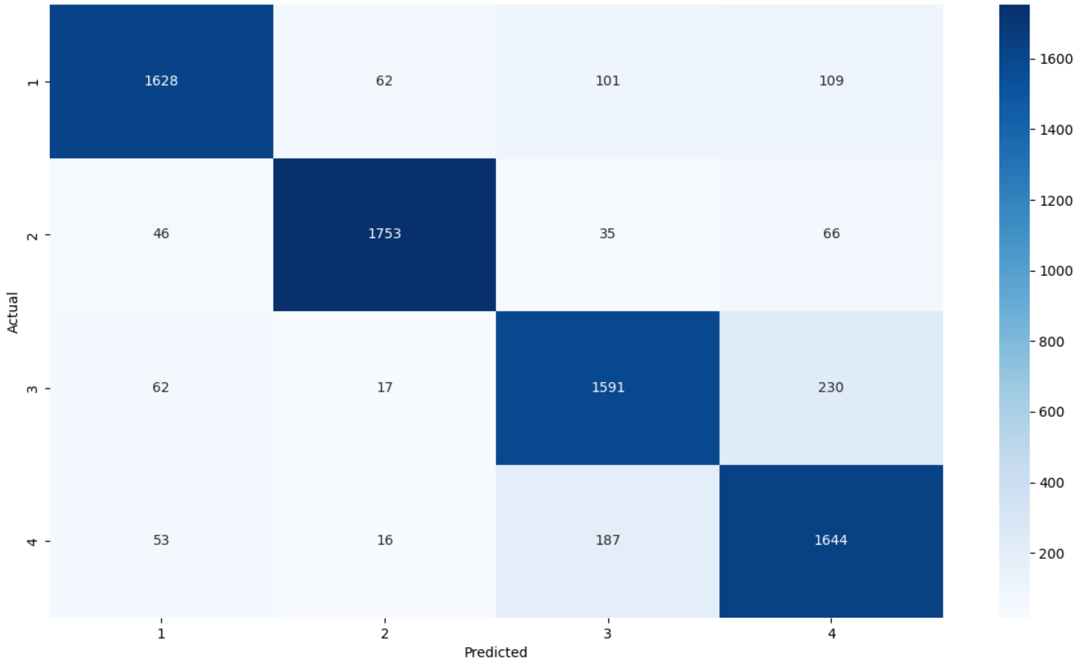
CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =10

- The training confusion matrix likely shows a predominance of true positives in each class, suggesting the model's proficiency in learning from the data.
- Given the extensive context provided by the larger window size, my SVD model has developed rich word embeddings, which usually results in a high accuracy during the training phase.



<u>CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =10</u>

- In the testing set, if the true positives are high and comparable to the training set, it signifies that the SVD model with a window size of 10 generalizes well.



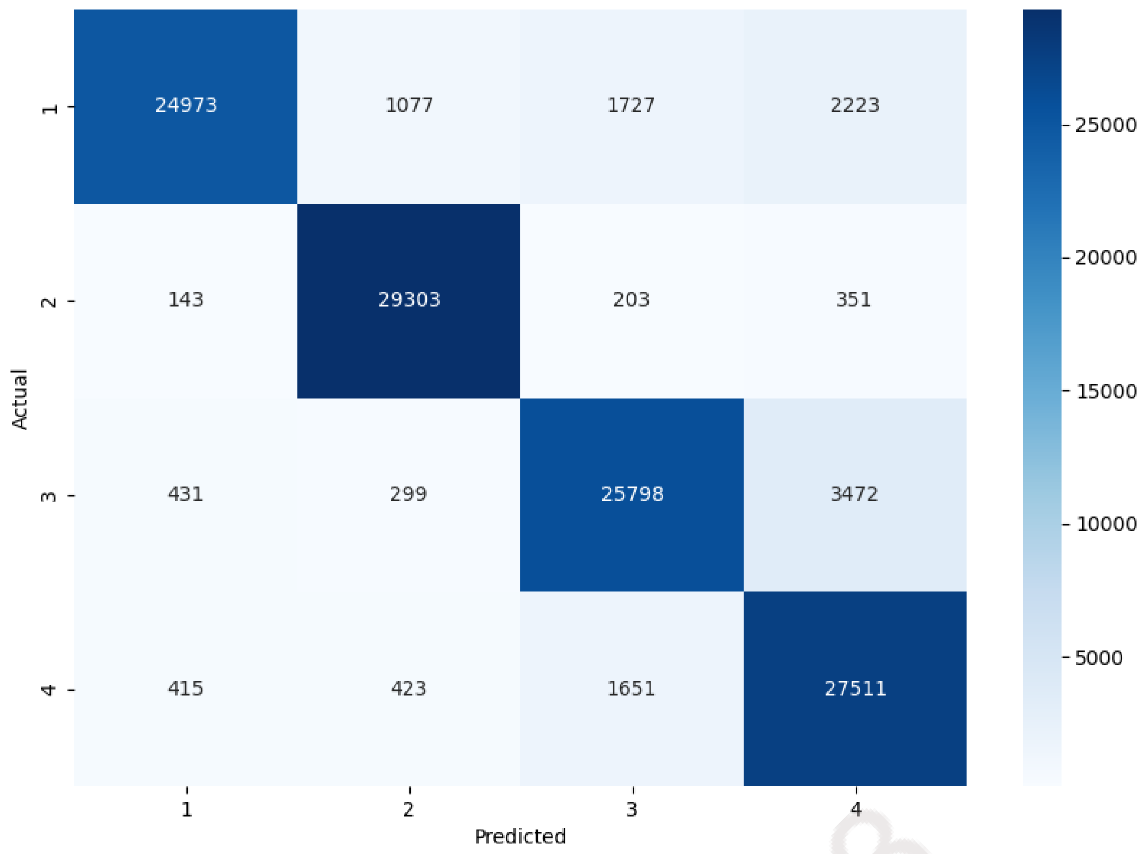
ANALYSIS WHEN WINDOW SIZE = 5

ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =5

```
Test Accuracy: 86.84%
Training Set Performance:
Accuracy: 0.90
F1 Score: 0.90
Precision: 0.90
Recall: 0.90

Test Set Performance:
Accuracy: 0.87
F1 Score: 0.87
Precision: 0.87
Recall: 0.87
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERAB
```

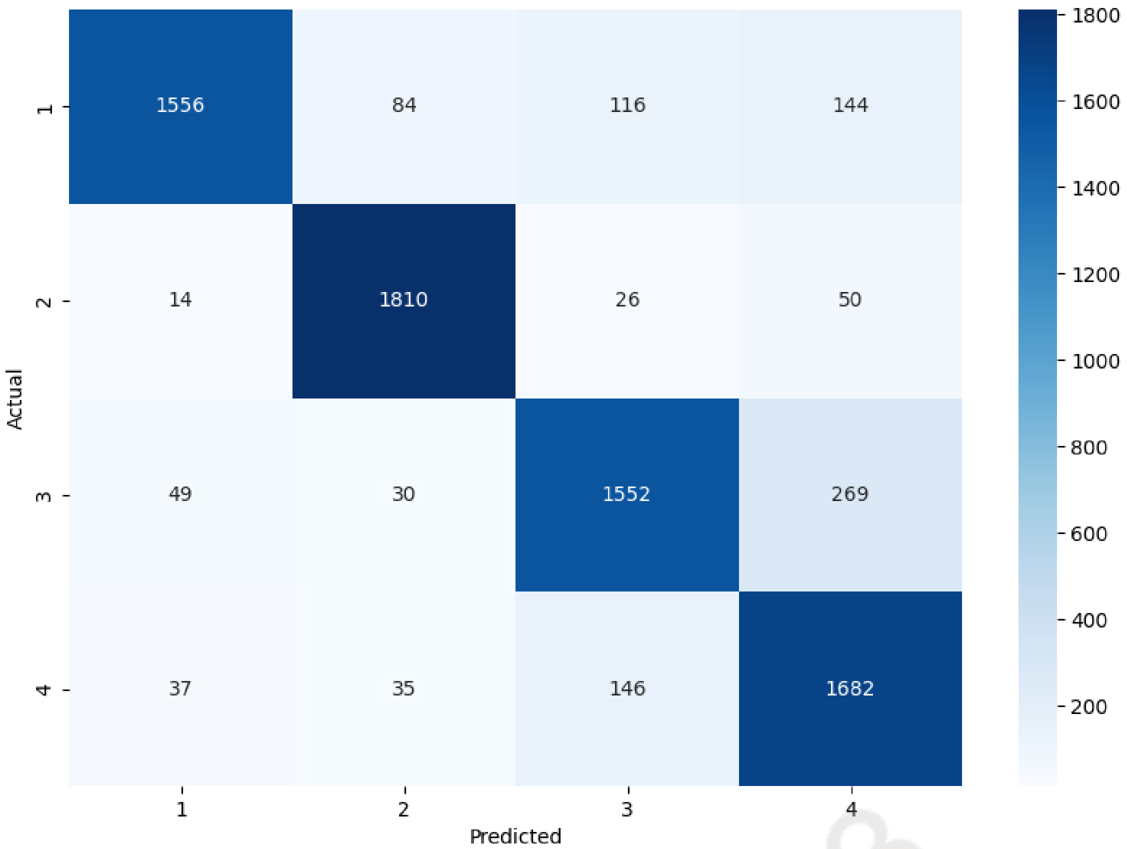
- Metrics for the window size of 5 show commendable results, with the training set achieving 0.90 across accuracy, F1 score, precision, and recall, while the testing set scores are closely aligned at 0.87.
- These metrics illustrate that the model maintains a high degree of reliability when processing unseen data.

<u>CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =5</u>



- Similarly, the training set's confusion matrix shows that the model has learnt well with high true positive rates (24973, 29303, 25798, 27511 for classes 1 to 4 respectively).

<u>CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =5</u>



- The training confusion matrix for a window size of 5 indicates strong performance, with a large number of true positives (1556, 1810, 1552, 1682 for classes 1 to 4 respectively) and relatively fewer misclassifications.

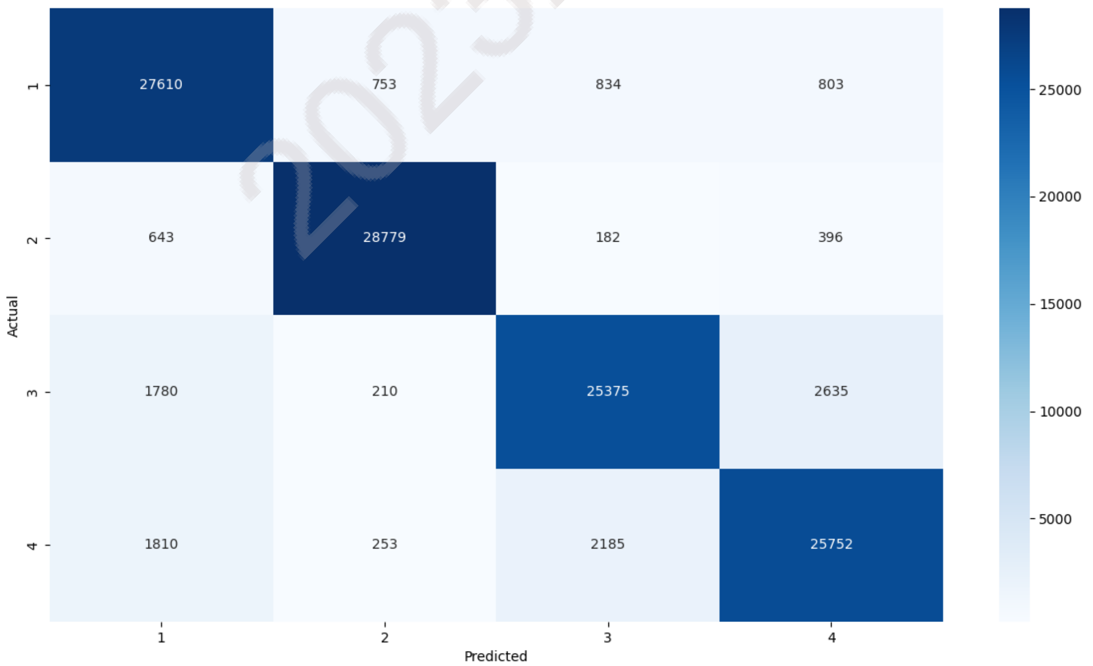
ANALYSIS WHEN WINDOW SIZE = 2

ACCURACY, RECALL AND F1 VALUES WHEN WINDOW SIZE =2

```
Test Accuracy: 86.61%
Training Set Performance:
Accuracy: 0.90
F1 Score: 0.90
Precision: 0.90
Recall: 0.90

Test Set Performance:
Accuracy: 0.87
F1 Score: 0.87
Precision: 0.87
Recall: 0.87
(myenv) PS C:\Users\nithi\OneDrive\Desktop\IIIT HYDERAB
```

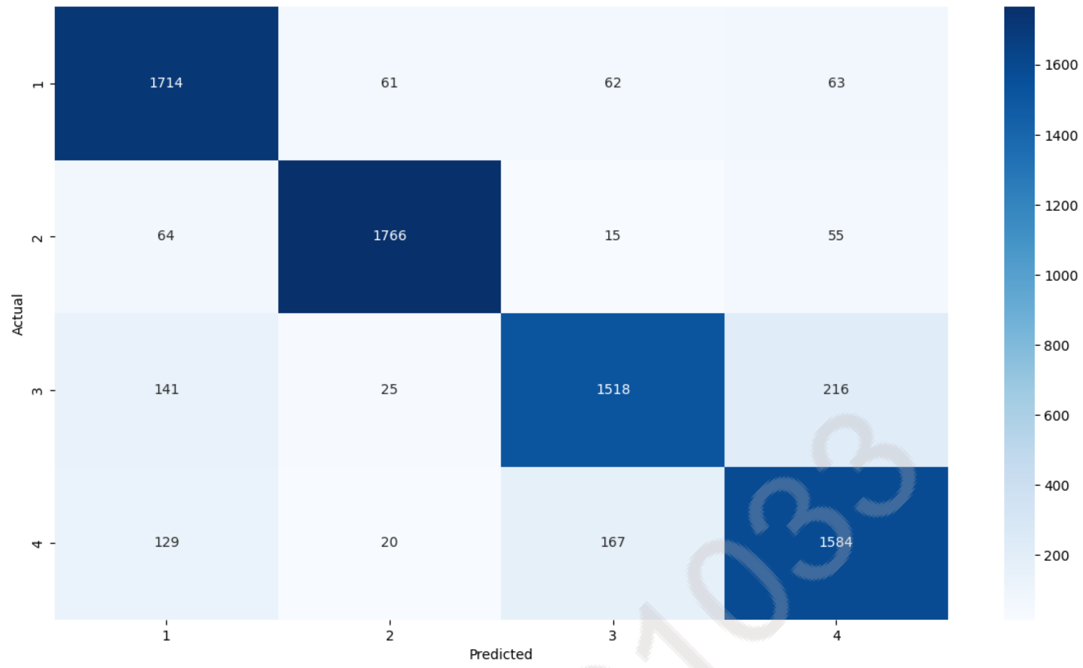
- The performance metrics   shows the model's effectiveness with an accuracy of 0.90 and an F1 score of 0.90 on the training data and an accuracy and F1 score of 0.87 on the testing data.
- These results indicate that the model retains a good level of predictive ability across both datasets, even when working with less context.

<u>CONFUSION MATRIX OF TRAINING SET WHEN WINDOW SIZE =2</u>



- For the training data, we can see a strong diagonal presence (27610, 28779, 25375, 27552 for classes 1 to 4), shows us that the model learns well from the training set.
- However, as with the testing data, the higher number of misclassifications compared to larger context windows points towards the model's difficulty in capturing more complex patterns due to the reduced context.

<u>CONFUSION MATRIX OF TESTING SET WHEN WINDOW SIZE =2</u>



- The testing confusion matrix for the window size of 2 shows a high number of correct predictions (1714, 1766, 1518, 1584 for classes 1 to 4, respectively), indicating that the model can accurately predict the test data.
- The off-diagonal numbers show some confusion between classes, **which is due to the limited contextual information available for each target word.**

<u>Comparative Analysis : SVD vs. Word2Vec (with Negative Sampling)</u>

- it is evident that **the skip gram with varied window sizes exhibits a strong capability in capturing the context of words.**
- **Larger window sizes up to a threshold generally yielded better performance, likely due to the increased contextual information** available for each target word, which is in line with the core design of Word2Vec.

- Word2Vec's **Skip-Gram model, particularly with larger context windows,** tends to perform better at capturing a wider range of lexical relationships, making it more adept at predicting word associations within broader contexts. **But at the same time, increasing the window size beyond a threshold is also deteriorating the performance.**
- Negative sampling which we added here helped the model to distinguish target words **from noise, thus reinforcing the word's context. I**t reduced the computational burden that arises from the need to update the weights for all words in the vocabulary for each training instance.
- Conversely, the results from the SVD approach, displays a slight underperformance in comparison to Word2Vec.
- SVD, which is a matrix factorization method might not capture word context as effectively as Word2Vec.
- This is due to SVD's inherent design, which compresses words into a dense representation but potentially loses some context.

Shortcomings

Shortcomings of SVD

- **Context Insensitivity**: SVD does not inherently account for the order of words, potentially leading to a loss in context that are critical for understanding context.
- **Scalability Issues**: SVD can be computationally expensive, especially with large vocabularies and documents, due to the need to decompose large matrices.
- **Static Embeddings**: The embeddings generated through SVD are static and do not adjust to new words is highly dependent on the size and quality of the training data. Insufficient or biased data can significantly affect the embeddings' quality.

Shortcomings of Word2Vec

- **Training Data Dependency**: Word2Vec's performance is highly dependent on the size and quality of the training data. Insufficient or biased data can significantly affect the embeddings' quality.
- **Computational Complexity**: Although more efficient than SVD, Word2Vec's Skip-Gram model with larger windows can **still be computationally demanding and it was compute heavy for me during training.**

HYPER PARAMETERS USED

- negative samples - 5
- window sizes - 2,5,10
- optimizer - Adam

Why i chose context window sizes as 2,5,10?

- **The smallest window size, 2**, was chosen to prioritize immediate context, emphasizing word pairs that are directly adjacent.

- **A window size of 5** represents a balanced approach, providing a broader context than size 2 but without the computational intensity of larger sizes. It captures not only immediate word interactions but also allows for the influence of surrounding words.

- **The largest window, size 10**, was chosen to explore the model's performance when a wide linguistic context is considered. It helps in understanding the thematic and topical associations between words that occur over longer distances within text

*****************END OF REPORT**********************