Multi-label Bibtex data Classification with SVM: Detailed Assignment Report

Introduction

This report elaborates on a multi-label classification task using Support Vector Machines (SVM) on the Bibtex dataset. The aim is to effectively predict multiple labels per instance, emphasizing the evaluation through metrics such as precision@5, accuracy, precision, recall, and F1 score, providing a detailed insight into model performance concerning data characteristics.

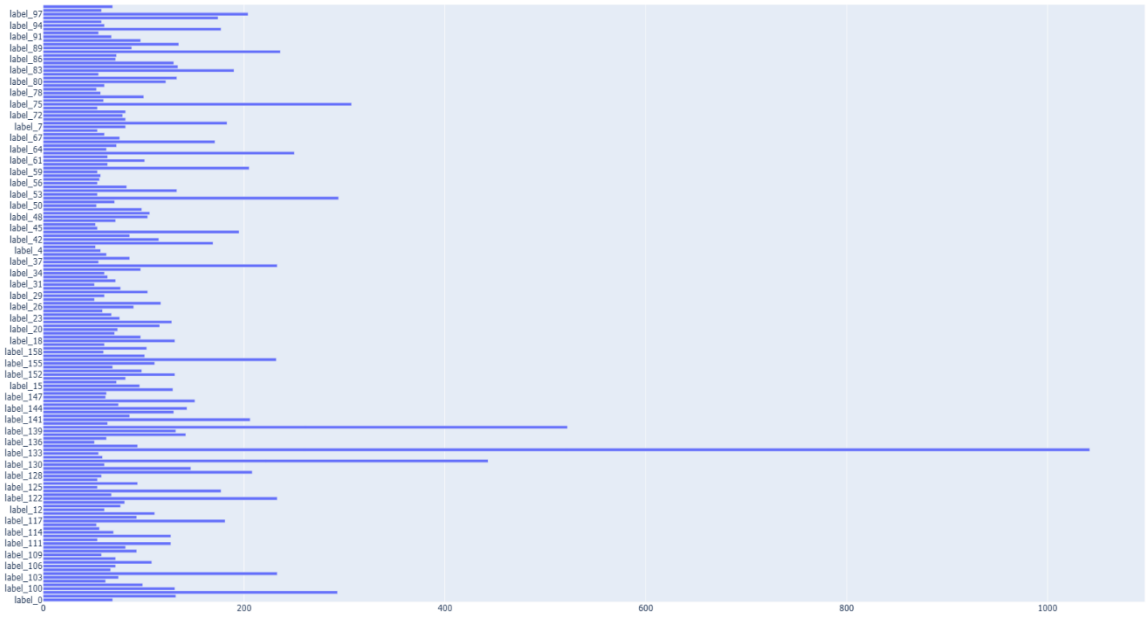
Dataset Analysis

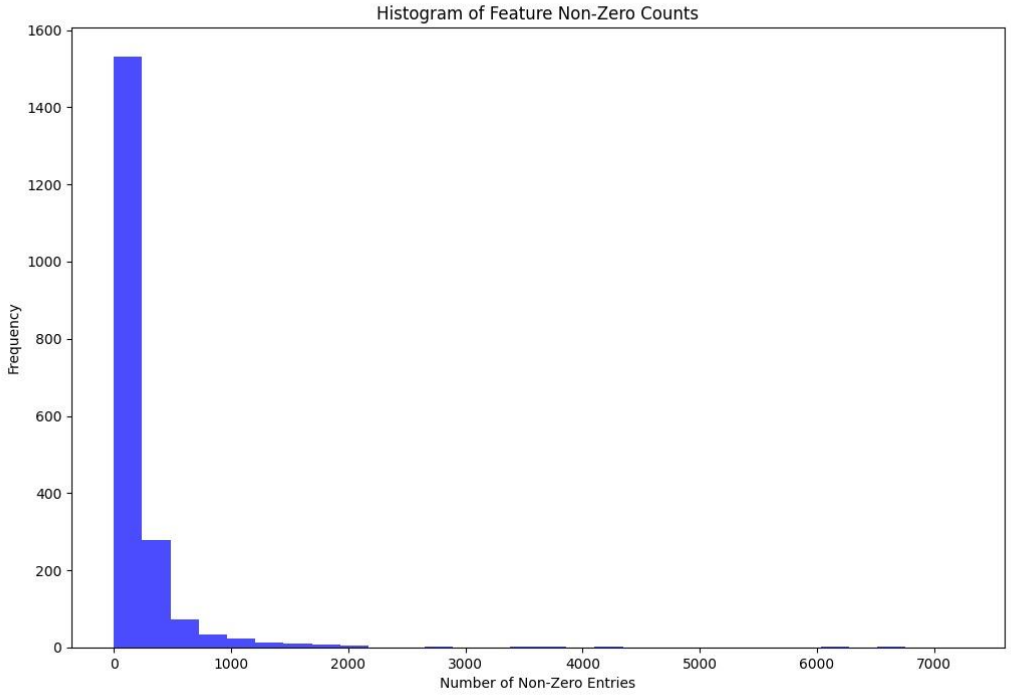
The Bibtex dataset contains multi-labeled data, presenting a typical sparse and high-dimensional feature space challenge inherent in classification tasks. The analysis of the dataset brought forth several critical insights, as outlined below:

Feature Sparsity and Label Frequency

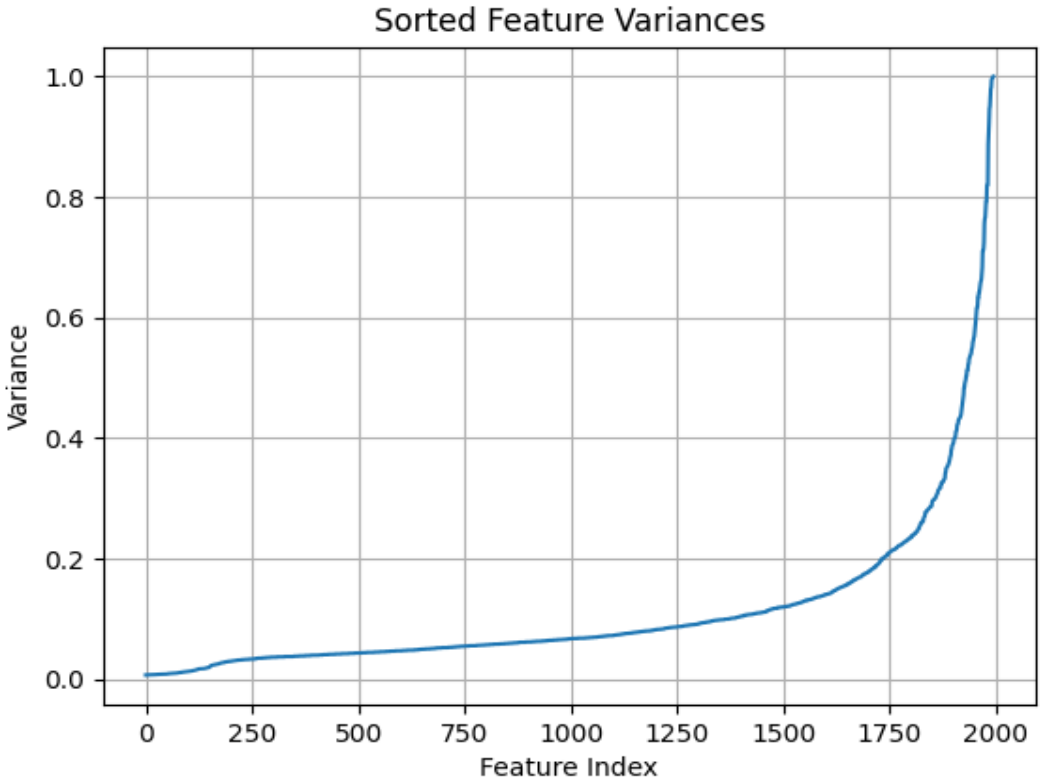
- Histogram of Label Frequency: The histogram of label frequency reveals a skewed distribution with a significant variance in label occurrence. Some labels are persistent, suggesting common topics or themes, while others appear infrequently, which poses challenges in achieving high predictive accuracy for these rare labels. This skewness can lead to biased models towards more common labels, potentially underfitting the less frequent but equally important categories.

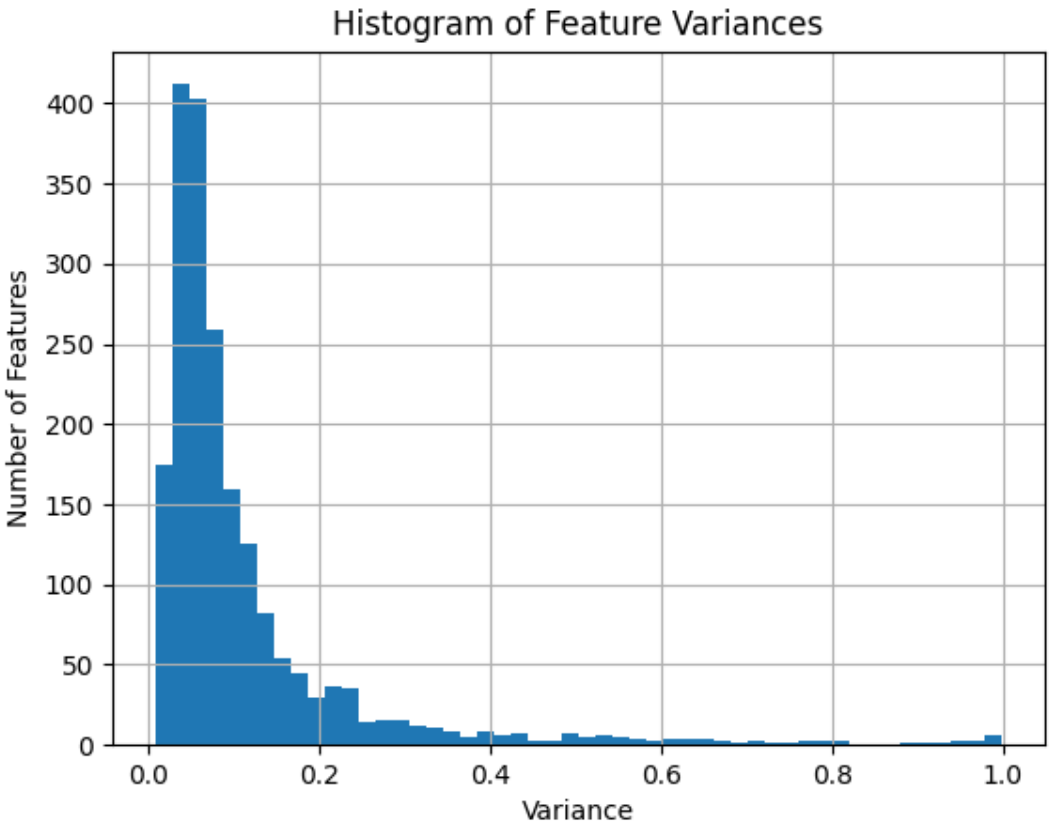
Non-Zero Feature Instance Counts: Most features have low non-zero counts, indicating a sparse matrix where many features are inactive across multiple samples. This sparsity challenges the model's learning process, as many features provide little information for distinguishing between classes.


label_97
label_94
label_91
label_89
label_86
label_83
label_80
label_78
label_75
label_72
label_7
label_67
label_64
label_61
label_59
label_56
label_53
label_50
label_48
label_45
label_42
label_4
label_37
label_34
label_31
label_29
label_26
label_23
label_20
label_18
label_158
label_155
label_152
label_15
label_147
label_144
label_141
label_139
label_136
label_133
label_130
label_128
label_125
label_122
label_12
label_117
label_114
label_111
label_109
label_106
label_103
label_100
label_0
0
200
400
600
800
1000


Histogram of Feature Non-Zero Counts
Frequency
Number of Non-Zero Entries
0
200
400
600
800
1000
1200
1400
1600
0
1000
2000
3000
4000
5000
6000
7000

- Feature Variances: The analysis of feature variances shows that many features have low variance, suggesting they do not contribute significantly to the model's decision-making process. Based on the findings, features with a variance below a threshold of 0.05 provide minimal information and could be candidates for removal to reduce the model complexity without sacrificing performance.





Methodology

Data Preprocessing

- *Text to DataFrame Conversion*: Conversion was performed to transform text data into a structured data frame, facilitating further manipulations. Labels were binarized using `*MultiLabelBinarizer*`, making them suitable for the SVM training process.

- *Feature Selection*: `*VarianceThreshold*` was applied to remove features with low variance, thus focusing the model training on more significant attributes.

- *Dimensionality Reduction*: `*PCA*` was used to reduce dimensionality for transformed features of degree 2 to make it computationally viable. The model could only capture a maximum of 40% variance for the transformed features.

Model Training and Evaluation

- *Model Configuration*: The `*OneVsRestClassifier*` wrapper around the SVM was used to adapt the binary classification method to multi-label tasks. Different configurations of SVM (linear and polynomial kernels) and regularization strengths (C values) were tested to find the optimal setup.

- *Evaluation Framework*: The model was evaluated on various metrics to understand its effectiveness from different perspectives, including:

1. Accuracy: This measures the overall correctness of the model, i.e., the ratio of accurate predictions (both true positives and true negatives) overall predictions.
2. Hamming Loss: It represents the fraction of incorrectly predicted labels.
3. Precision: This indicates the accuracy of optimistic predictions, formulated as the ratio of true positives to the sum of true and false positives.
4. Recall: This measures the ability of the classifier to find all the positive samples, calculated as the ratio of true positives to the sum of true positives and false negatives.
5. F1 Score: This is the harmonic mean of precision and recall, providing a single metric to assess accuracy that balances precision and recall.
6. Precision@5: This metric evaluates precision by considering only the top 5 scores predicted by the model, reflecting its effectiveness in identifying the most relevant labels.

Results

Model Performance Metrics

The results showed varied performance across different model configurations, significantly influenced by choice of kernel, the number of principal components, and the C parameter. *Due to*

computational resource constraints, the polynomial models were able to run with features that explain only 40% of the variance in the data:

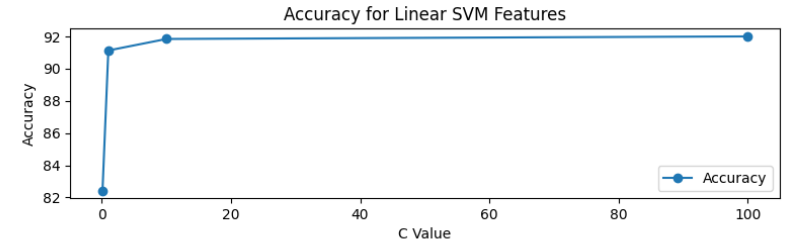
Linear SVM using `SVC` without dimensionality reduction results:

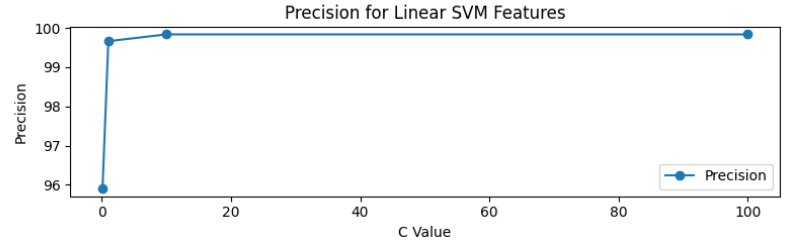
- Accuracy of Linear SVM with C = 0.1: 82.42544731610339
- Hamming loss of Linear SVM with C = 0.1: 0.14429148380159296
- Precision of Linear SVM with C = 0.1: 95.90307485135897
- Recall of Linear SVM with C = 0.1: 96.2323132609411
- F1 Score of Linear SVM with C = 0.1: 95.68546466843797
- Precision@5 with C = 0.1: 99.72081867767754

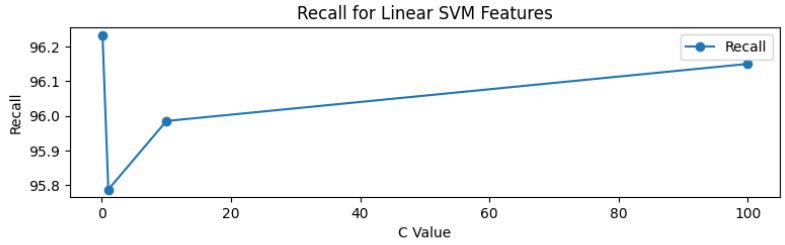
- Accuracy of Linear SVM with C = 1: 91.13320079522863
- Hamming loss of Linear SVM with C = 1: 0.06926991510059143
- Precision of Linear SVM with C = 1: 99.66441811597994
- Recall of Linear SVM with C = 1: 95.7880881869036
- F1 Score of Linear SVM with C = 1: 97.6036854708812
- Precision@5 with C = 1: 99.93090977941873

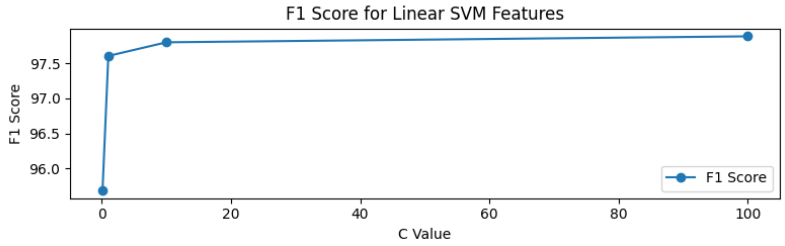
- Accuracy of Linear SVM with C = 10: 91.84890656063618
- Hamming loss of Linear SVM with C = 10: 0.0635182615001813
- Precision of Linear SVM with C = 10: 99.83824266661095
- Recall of Linear SVM with C = 10: 95.98552155314249
- F1 Score of Linear SVM with C = 10: 97.79815226708766
- Precision@5 with C = 10: 99.93081510934394

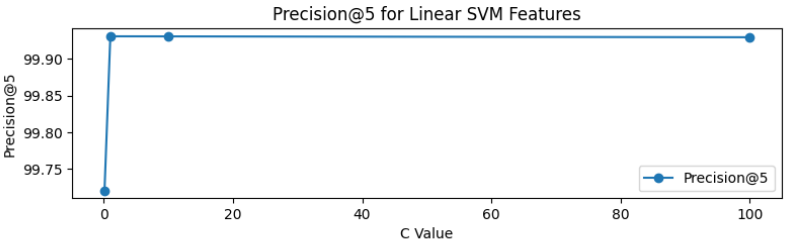
- Accuracy of Linear SVM with C = 100: 92.0079522862823
- Hamming loss of Linear SVM with C = 100: 0.0610175425434812 54
- Precision of Linear SVM with C = 100: 99.83824266661095
- Recall of Linear SVM with C = 100: 96.15004935834156
- F1 Score of Linear SVM with C = 100: 97.88318993212832
- Precision@5 with C = 100: 99.92961352762546

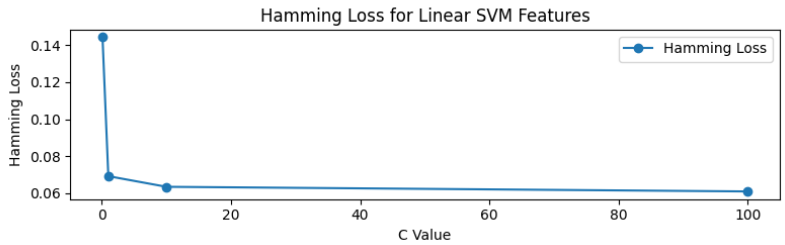

Accuracy for Linear SVM Features
Accuracy
C Value
Accuracy


Precision for Linear SVM Features
Precision
C Value
Precision


Recall for Linear SVM Features
Recall
C Value
Recall


F1 Score for Linear SVM Features
F1 Score
C Value
F1 Score


Precision@5 for Linear SVM Features
Precision@5
C Value
Precision@5


Hamming Loss for Linear SVM Features
Hamming Loss
C Value
Hamming Loss

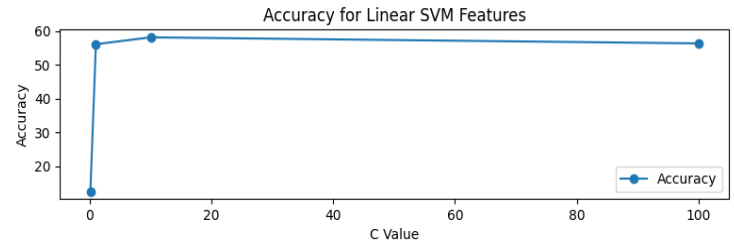
Linear SVM using `SVC` with dimensionality reduction (with features capturing only 40% variance) results:

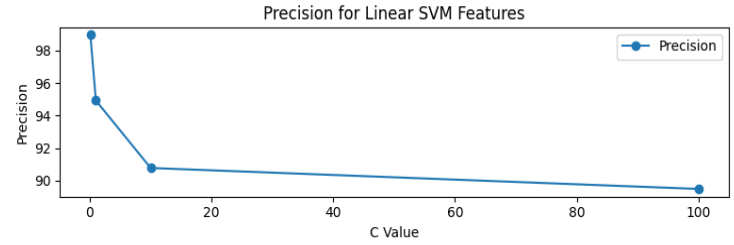
- Accuracy of Linear SVM with C = 0.1: 12.604373757455267
- Hamming loss of Linear SVM with C = 0.1: 1.2566112757417758
- Precision of Linear SVM with C = 0.1: 98.95806872841518
- Recall of Linear SVM with C = 0.1: 17.900625205659757
- F1 Score of Linear SVM with C = 0.1: 20.313805651605538
- Precision@5 with C = 0.1: 95.0081723378945

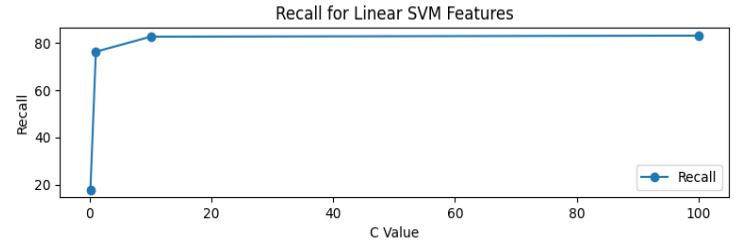
- Accuracy of Linear SVM with C = 1: 56.063618290258454
- Hamming loss of Linear SVM with C = 1: 0.4233717193693187
- Precision of Linear SVM with C = 1: 94.92242927652936
- Recall of Linear SVM with C = 1: 76.40671273445211
- F1 Score of Linear SVM with C = 1: 83.88447602034425
- Precision@5 with C = 1: 95.99476299196259

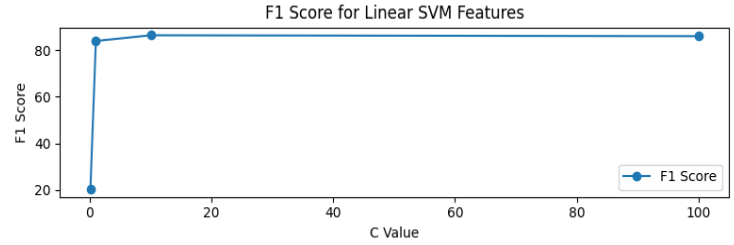
- Accuracy of Linear SVM with C = 10: 58.09145129224652
- Hamming loss of Linear SVM with C = 10: 0.39211273241056804
- Precision of Linear SVM with C = 10: 90.77568189759528
- Recall of Linear SVM with C = 10: 82.77393879565646
- F1 Score of Linear SVM with C = 10: 86.32681654876802
- Precision@5 with C = 10: 94.77958861662049

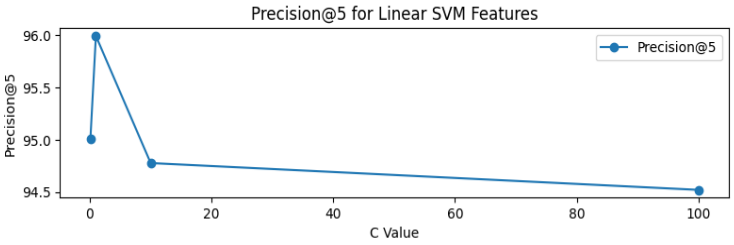
- Accuracy of Linear SVM with C = 100: 56.30218687872763
- Hamming loss of Linear SVM with C = 100: 0.4101179088988084
- Precision of Linear SVM with C = 100: 89.48604891511575
- Recall of Linear SVM with C = 100: 83.21816386969397
- F1 Score of Linear SVM with C = 100: 85.93661063850739
- Precision@5 with C = 100: 94.523094764205

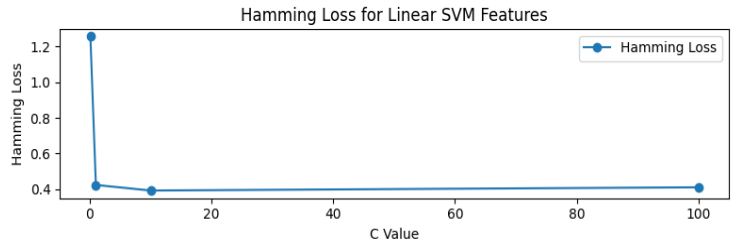

Accuracy for Linear SVM Features
Accuracy
C Value
Accuracy


Precision for Linear SVM Features
Precision
C Value
Precision


Recall for Linear SVM Features
Recall
C Value
Recall


F1 Score for Linear SVM Features
F1 Score
C Value
F1 Score


Precision@5 for Linear SVM Features
Precision@5
C Value
Precision@5


Hamming Loss for Linear SVM Features
Hamming Loss
C Value
Hamming Loss

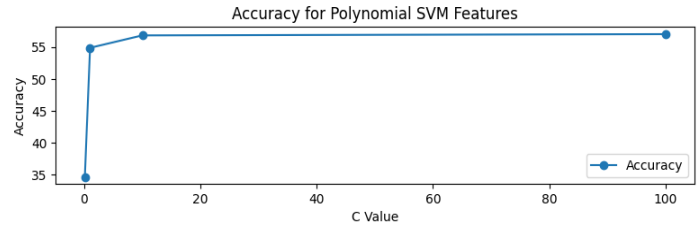
Polynomial Transformed features with Linear SVM using `SVC` with dimensionality reduction (with features capturing only 40% variance) results:

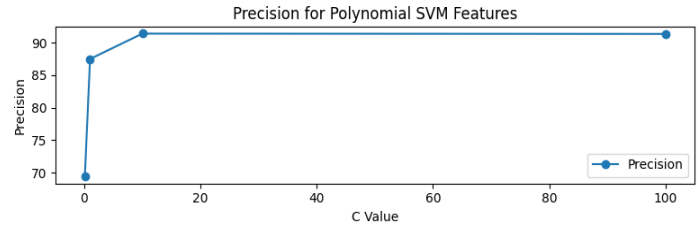
- Accuracy of POLY SVM with C = 0.1: 34.67196819085487
- Hamming loss of POLY SVM with C = 0.1: 0.9042599747427386
- Precision of POLY SVM with C = 0.1: 69.43375588188 94
- Recall of POLY SVM with C = 0.1: 90.83580125041132
- F1 Score of POLY SVM with C = 0.1: 77.24178199448075
- Precision@5 with C = 0.1: 92.46206381608938

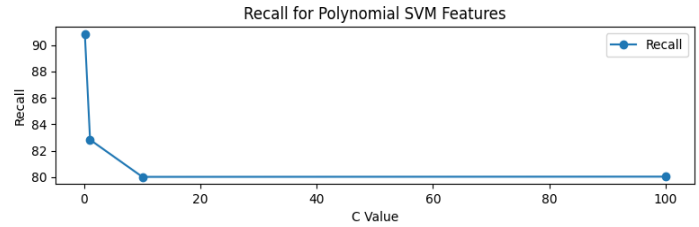
- Accuracy of POLY SVM with C = 1: 54.91053677932406
- Hamming loss of POLY SVM with C = 1: 0.4528802030583793
- Precision of POLY SVM with C = 1: 87.50441381225113
- Recall of POLY SVM with C = 1: 82.82329713721619
- F1 Score of POLY SVM with C = 1: 84.61451951030051
- Precision@5 with C = 1: 94.1831710650267

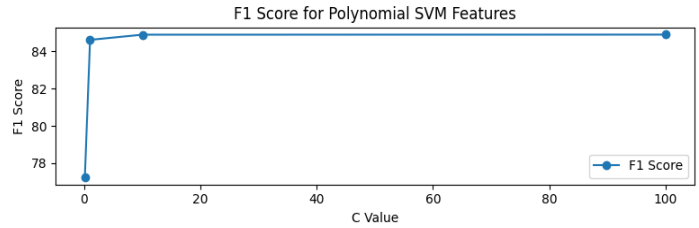
- Accuracy of POLY SVM with C = 10: 56.85884691848907
- Hamming loss of POLY SVM with C = 10: 0.4201207847256086
- Precision of POLY SVM with C = 10: 91.40790217900 72
- Recall of POLY SVM with C = 10: 80.00987166831194
- F1 Score of POLY SVM with C = 10: 84.89280630257718
- Precision@5 with C = 10: 94.32124910951863

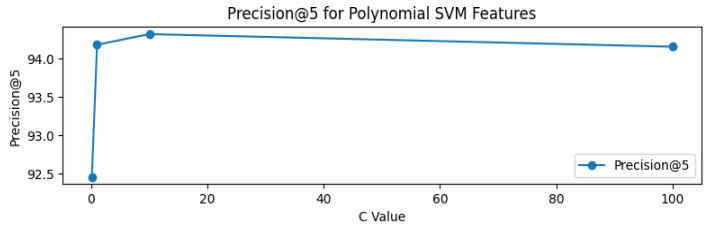
- Accuracy of POLY SVM with C = 100: 57.057654075546715
- Hamming loss of POLY SVM with C = 100: 0.42037085662127865
- Precision of POLY SVM with C = 100: 91.360857158278
- Recall of POLY SVM with C = 100: 80.02632444883186
- F1 Score of POLY SVM with C = 100: 84.89633356424335
- Precision@5 with C = 100: 94.15708942779099

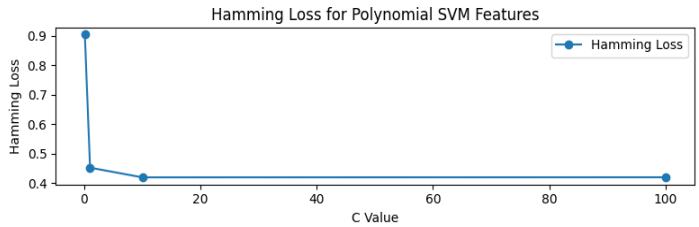

Accuracy for Polynomial SVM Features
Accuracy
C Value
Accuracy


Precision for Polynomial SVM Features
Precision
C Value
Precision


Recall for Polynomial SVM Features
Recall
C Value
Recall


F1 Score for Polynomial SVM Features
F1 Score
C Value
F1 Score


Precision@5 for Polynomial SVM Features
Precision@5
C Value
Precision@5


Hamming Loss for Polynomial SVM Features
Hamming Loss
C Value
Hamming Loss

Linear SVM Result Analysis

1. Improvement with Increased C:

- As `C` increases from 0.1 to 100, there's a consistent improvement in most metrics. In particular, accuracy, F1 score, and precision improved notably.

- The increase in `C` generally decreases the strength of regularization, allowing the model to fit more closely to the training data, which might be beneficial if the model was initially underfitting.

2. Diminishing Returns:

- The improvements in performance metrics from `C` = 10 to `C` = 100 are less pronounced compared to earlier increments. This suggests approaching a point of diminishing returns where increasing `C` further may not yield significant benefits and could lead to overfitting, especially noticeable if tested on a completely independent test set.

3. Hamming Loss:

- The hamming loss decreases as `C` increases, which is a positive sign. Lower hamming loss means the model makes fewer mistakes across the multiple labels it is predicting.

4. Precision and Recall:

- The precision is exceptionally high, particularly at `C` = 1, 10, 100, suggesting the model is reliable when predicting a positive label. The minor improvements in recall across different values of `C` indicate that the model is gradually getting better at identifying all relevant labels without increasing false positives.

5. Precision@5:

- Precision@5 remains high across all configurations, indicating that the top 5 predicted labels by the model are very likely to be true positives. This high value is crucial for applications where the top few predictions are more important than the complete set of outputs.

Conclusion:

The model performs exceptionally well across all tested configurations, with powerful results in precision and precision@5. This indicates that the linear SVM is a robust choice for this multi-label classification task.

Polynomial SVM Result Analysis

1. Increasing Regularization Parameter (C):

- Almost all performance metrics show a notable increase as `C` increases from 0.1 to 100. This suggests that reducing regularization (allowing the model to fit more closely to the data) helps, particularly in a setting where the feature space does not fully capture the dataset's variance.

2. Low Accuracy:

- The accuracy starts relatively low at 34.67% and increases to 57.06% but remains modest. This indicates the limitations due to using only 40% of the variance. The model may not capture enough information to classify many instances accurately.

3. High Recall, Moderate Precision:

- The recall is significantly high, especially at lower `C` values, indicating that the model is good at identifying positive labels but at the cost of also misclassifying many negatives as positives (as evidenced by lower precision at `C` = 0.1).

- As `C` increases, precision improves considerably, suggesting better differentiation between classes at the cost of a slight drop in recall.

4. Improvement in F1 Score:

- The F1 score, which balances precision and recall, shows improvement as `C` increases, peaking at `C` = 1 and `C` = 10, which might be the best trade-off between avoiding overfitting while maintaining good model performance.

5. Precision@5:

- The precision@5 is very high across all `C` values, suggesting that the top 5 predictions are generally accurate. This is crucial for applications where the top few predictions matter the most, such as ranking or recommendation systems.

Conclusion:

The performance of the polynomial SVM on your dataset, with current computational constraints, shows reasonable effectiveness, especially in identifying the most relevant labels (as indicated by high Precision@5). However, the overall accuracy suggests significant room for improvement, likely hindered by the limited variance captured in the feature set.

Comparative Analysis

Comparing the performance of the standard linear SVM and the polynomial-transformed features of linear SVM, both with dimensionality reduction captured only 40% of the variance.

Linear SVM:

1. Accuracy:

- Starts very low at 12.60% for C=0.1, suggesting poor model fit at this regularization strength.

- Significantly improves as C increases, peaking at 58.09% for C=10, then slightly drops at C=100.

2. Hamming Loss:

- Reduces as C increases, indicating fewer incorrect label predictions with higher C values.

3. Precision and Recall:

- Precision starts high at 98.95% for C=0.1 but drops as C increases, suggesting a trade-off between avoiding overly confident incorrect predictions and missing out on correct predictions.

- Recall improves dramatically as C increases, indicating better coverage of positive labels.

4. F1 Score:

- Reflects the trade-offs, with improvements as C increases, peaking at C=10.

5. Precision@5:

- Remains relatively high across all values of C, suggesting that the model performs well when predicting the top five most probable labels.

Polynomial SVM (Transformed Features):

1. Accuracy:

- Considerably higher starting accuracy at 34.67% for C=0.1, suggesting that polynomial features offer better initial separability, even with limited variance.

- Increases with C, but the improvements taper off, indicating diminishing returns with increasing regularization strength.

2. Hamming Loss:

- Higher than the linear case at C=0.1, but rapidly improves as C increases.

3. Precision and Recall:

- Both are significantly lower than in the linear case at C=0.1 but improve rapidly with increasing C, suggesting that the model becomes more discerning with higher regularization.

4. F1 Score:

It also starts lower than the linear model at C=0.1 but exceeds all linear F1 scores as C increases, indicating a better overall balance between precision and recall in the polynomial model.

5. Precision@5:

- Although lower than the linear model at C=0.1, it improves and remains competitive across all C values.

Overall Comparison:

- Initial Model Fit: Polynomial features provide a better starting point for model performance even with the same variance captured, likely due to the enhanced separability from feature interactions.

- Impact of Regularization (C): Both models improve with increased regularization strength, but the polynomial model shows more consistent and substantial gains in recall and F1 scores. This might indicate that polynomial features respond better to less regularization in this dataset.

- Best Use of Computational Resources: Given the similar computational costs, polynomial features provide a better return

on investment regarding model performance metrics, especially in scenarios where recall and F1 score are critical.

Conclusion:

Increasing the variance captured or further tuning the regularization parameter might provide better results if computational resources allow. Despite the same limitation in variance capture, the polynomial model consistently outperforms the linear model in most metrics, particularly in ensuring robust recall and F1 scores, making it a preferable choice for complex data structures like those in the Bibtex dataset.

Discussion

The comprehensive analysis suggests that the dataset's complexity, characterized by high dimensionality and sparsity, necessitates sophisticated modeling techniques. Polynomial SVM's superior performance can be attributed to its ability to model intricate patterns not capturable by the linear SVM, especially in a multi-label context where interactions between labels are complex.

Final Thoughts:

This extensive investigation into various SVM configurations for the complex multi-label classification task has underscored the critical roles of feature engineering, model selection, and parameter optimization. Future endeavors could explore advanced machine learning techniques to enhance model performance further:

Deep Learning Approaches: Techniques such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) could be leveraged to capture the sequential patterns in text data more effectively.

Advanced Ensemble Methods: Combining multiple models in an ensemble method, such as random forests or boosted trees, might improve prediction accuracy by reducing the likelihood of overfitting the dataset's noisy or less informative features.

Hybrid Models: Integrating SVM with neural network architectures might balance the robustness of SVMs and the learning capacity of deep networks, especially for handling high-dimensional and sparse data like the Bibtex dataset.

References:

- Multiclass Classification Using Support Vector Machines | Baeldung on Computer Science
- python 2.7 - Scikit-learn 0.15.2 - OneVsRestClassifier not works due to predict_proba not available - Stack Overflow