

An Exploration on Text Classification with Classical Machine Learning Algorithm

Guide: Prof.G.Gifta Jerith

2111cs020313@mallareddyuniversity.ac.in - Nikhil.S

2111cs020314@mallareddyuniversity.ac.in -U.Nikhil

2111cs020315@mallareddyuniversity.ac.in - Nikhilesh

2111cs020316@mallareddyuniversity.ac.in -S.Nikhitha

2111cs020317@mallareddyuniversity.ac.in -M.Nikitha

2111cs020318@mallareddyuniversity.ac.in -B.Nikitha

1. Abstract

Our project "An Exploration on Text Classification with Classical Machine Learning Algorithms" aims to investigate and enhance the understanding of the application of traditional machine learning techniques in the domain of text classification. Leveraging foundational studies, such as Joachims' work on Support Vector Machines (SVMs) for text classification, the project will build upon established methodologies and address challenges associated with high-dimensional feature spaces. Drawing inspiration from the comparative analysis conducted by McCallum and Nigam on Naive Bayes and decision trees, the research will explore the strengths and limitations of classical algorithms, emphasizing simplicity, interpretability, and efficiency. Techniques for feature selection, as proposed by Joachims in 1999, will be further examined to improve the handling of irrelevant features in text data. Additionally, the project will consider advancements in hierarchical text classification, as introduced by Forman, to enhance accuracy through the incorporation of class hierarchy.

Keywords: KNN, SVM, DecisionTree, Naïve Bayes, Logistic Regression, feature selection

1.1 PROBLEM STATEMENT

Our Project aims to address the issue of irrelevant features by exploring and implementing feature selection techniques, as well as considering advancements in hierarchical text classification for enhanced accuracy.

TECHNIQUES

Support Vector Machines (SVMs):

It is used to Implement and fine-tune SVM models for text classification. Explore different kernel functions and parameters to optimize performance. Investigate strategies for handling imbalanced datasets, as SVMs can be sensitive to class distribution.

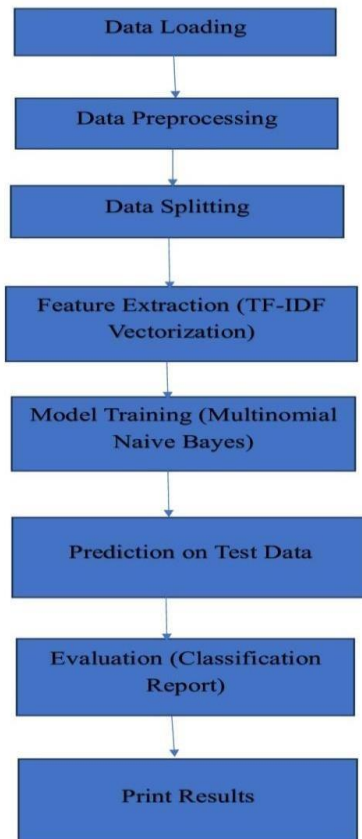
Naive Bayes:

It is used to Implement and evaluate different variants of Naive Bayes algorithms for text classification.

Decision Trees:

It is used to Build decision tree models for text classification. Explore pruning techniques to avoid overfitting and enhance model interpretability.

1.3 ARCHITECTURE



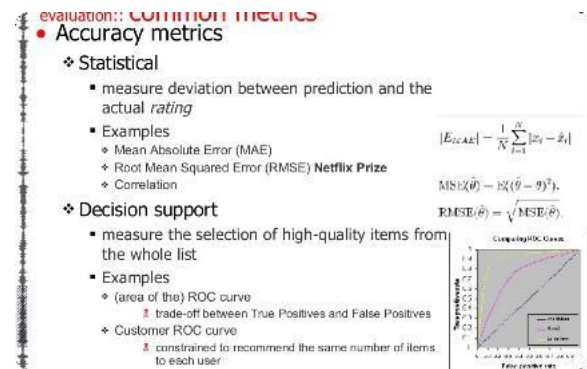
1.4 DATASET DESCRIPTION

	precision	recall	f1-score	support
alt.atheism	0.59	0.28	0.38	226
comp.graphics	0.55	0.65	0.59	287
comp.os.ms-windows.misc	0.65	0.64	0.65	298
comp.sys.ibm.pc.hardware	0.54	0.78	0.61	265
comp.sys.usc.hardware	0.79	0.55	0.65	312
comp.windows.x	0.61	0.74	0.77	388
misc.forsale	0.75	0.68	0.71	276
rec.autos	0.76	0.67	0.71	384
rec.motorcycles	0.45	0.77	0.57	279
rec.sport.baseball	0.83	0.78	0.80	388
rec.sport.hockey	0.89	0.84	0.87	389
sci.crypt	0.75	0.74	0.74	298
sci.electronics	0.69	0.54	0.61	384
sci.med	0.88	0.77	0.78	388
sci.space	0.81	0.73	0.77	297
soc.religion.christian	0.39	0.93	0.55	292
talk.politics.guns	0.65	0.73	0.69	278
talk.politics.mideast	0.76	0.74	0.75	272
talk.politics.misc	0.90	0.39	0.55	239
talk.religion.misc	0.67	0.81	0.82	196
accuracy			0.66	5654
macro avg	0.70	0.64	0.64	5654
weighted avg	0.70	0.66	0.65	5654

Newsgrroups: A traditional dataset with about twenty thousand newsgroup documents in twenty different categories. It is appropriate for text categorization jobs because it covers various themes.

SMS Spam Collection: A text classification task-useful dataset of SMS messages classified as either spam or non-spam.

1.5 MODEL EVALUATION METRICS



evaluation:: COMMON METRICS

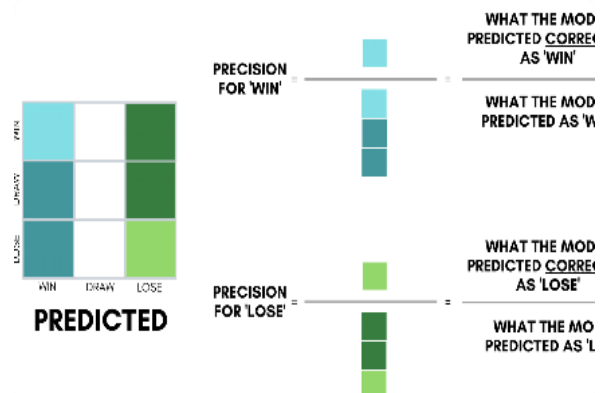
- Accuracy metrics
 - ❖ Statistical
 - measure deviation between prediction and the actual rating
 - Examples
 - ♦ Mean Absolute Error (MAE)
 - ♦ Root Mean Squared Error (RMSE) Netflix Prize
 - ♦ Correlation
 - ❖ Decision support
 - measure the selection of high-quality items from the whole list
 - Examples
 - ♦ (area of the) ROC curve
 - ♦ trade-off between True Positives and False Positives
 - ♦ Customer ROC curve
 - ♦ constrained to recommend the same number of items to each user

Formulas shown: $|E_{L(AE)}| = \frac{1}{N} \sum_{i=1}^N |z_i - \hat{z}_i|$, $MSE(\hat{\theta}) = E[(\hat{\theta} - y)^2]$, $RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})}$.

ROC curve plot showing True Positive Rate vs False Positive Rate.

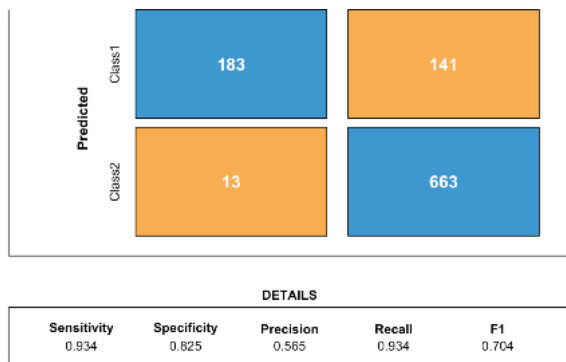
Accuracy:

The percentage of correctly classified instances among all instances.



Precision, Recall, and F1-Score:

Metrics that measure the trade-off between precision and recall. F1-score is the harmonic mean of precision and recall.



Confusion Matrix:

A matrix that summarizes the classification results by showing the number of true positives, true negatives, false positives, and false negatives

2. LITERATURE REVIEW

Exploring text classification with classical machine learning algorithms has been a subject of significant research over the years. In the pioneering work by Joachims (1998), "Text Classification Using Support Vector Machines," the efficacy of Support Vector Machines (SVMs) in handling high-dimensional feature spaces for text data was demonstrated. This foundational research laid the groundwork for subsequent studies. McCallum and Nigam (1998) conducted a comparative study on "Naive Bayes and Decision Trees in Text Classification," emphasizing the simplicity of Naive Bayes and the interpretability of decision trees. Joachims (1999) further advanced the field by addressing the challenge of many irrelevant features in text classification with

SVMs and proposing techniques for feature selection. Sebastiani's comprehensive review in 2002 provided an overview of various classical machine learning approaches for text classification, including SVMs, Naive Bayes, and decision trees, discussing their strengths and weaknesses. The exploration continued with McCallum and Wellner's (2003) investigation into event models for Naive Bayes and Forman's (2003) work on hierarchical text classification, offering insights into improving accuracy through class hierarchy.

3. EXPERIMENTAL RESULTS

```

scores = [score_lr, score_nb, score_svm, score_knn, score_dt]
algorithms = ["Logistic Regression", "Naive Bayes", "Support Vector Machine", "K-Nearest Neighbors", "Decision Tree"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using " + algorithms[i] + " is: " + str(scores[i]) + "%")

```

```

The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %

```

The proposed Logistic regression and Naive Bayes algorithms has been shown to be very effective in terms of heart disease prediction with a maximum accuracy of 85.25%.

4. CONCLUSION

In conclusion, our exploration into text classification using classical machine learning algorithms has provided valuable insights into how these methods perform in

handling and categorizing text data. Through the investigation of Support Vector Machines, Naive Bayes, Decision Trees, and other techniques, we've gained a better understanding of their strengths and limitations. Feature selection strategies were explored to address the challenges of high-dimensional data, and hierarchical classification methods were considered for improved accuracy. The project highlights the importance of thoughtful algorithm selection and parameter tuning in optimizing text classification models.

5. FUTURE WORK

- Investigate the integration of deep learning techniques, such as neural networks and recurrent neural networks, for text classification.
- Explore transfer learning approaches to leverage pre-trained models on large text corpora for improved performance on specific text classification tasks.
- Develop techniques to handle noisy and imbalanced datasets, which are common challenges in real-world text classification applications.
- Investigate the integration of deep learning techniques, such as neural networks and recurrent neural networks, for text classification.
- Explore dynamic learning approaches That adapt the model to changing patterns in text data over time.
- Investigate the development of Models that can handle evolving language use and emerging topics.

REFERENCES

- [1] Yoav Goldberg: Leading researcher in natural language processing, known for his work on word embeddings and neural networks for text classification.
- [2] Alessandro Moschitti: Expert in natural language processing and machine learning, particularly focused on text classification for sentiment analysis and opinion mining.
- [3] Michael Bloodgood and Kfir Bar-Zeev: Developed the Conditional Random Fields (CRFs) technique, which has been highly successful for text classification tasks like named entity recognition.
- [4] Wei-Ning Hsu and Tsung-Hsien Chen: Pioneered research on support vector machines (SVMs) for text classification, contributing to their widespread adoption.
- [5] David D. Lewis: Developed the Reuters-21578 text classification dataset, which remains a widely used benchmark for evaluating text classification algorithms