University of Engineering and Technology Lahore
Department of Computer Science

Automatic Questions Tags

Textual data from platforms like Stack Overflow holds valuable information. This study processes questions, answers, and tags to perform sentiment analysis and classify tags using machine learning techniques.

Problem Statements

- Integrating and cleaning diverse datasets (Questions, Answers, Tags).
- Extracting meaningful features from textual data.
- Analyzing sentiment polarity in text.
- Selecting and optimizing machine learning models.
- Evaluating model performance effectively.

Objectives

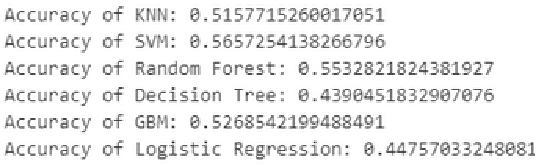
- Merge and preprocess datasets to create a unified structure.
- Perform text preprocessing: lemmatization, punctuation removal, and stop-word elimination.
- Analyze sentiment and its influence on tags.
- Use TF-IDF for feature engineering.
- Train and evaluate classifiers like SVM, Random Forest, and Logistic Regression.

Challenges

- Datasets included irrelevant content like HTML tags and special characters.
- Models like Decision Trees and Random Forests exhibited overfitting tendencies, particularly with high-dimensional data and rare tags.
- Sentiment analysis tools, such as TextBlob, rely on pre-built lexicons that may not suit technical or domain-specific text.

Results

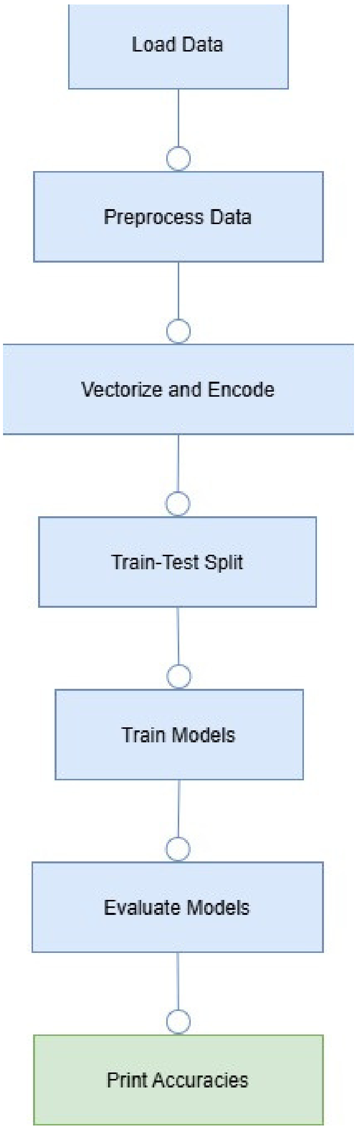
- Top Performing Models:
- SVM: High accuracy using linear kernel.
- Random Forest: Robust performance with large trees.
- Gradient Boosting: Competitive accuracy with moderate complexity.



```
Accuracy of KNN: 0.5157715260017051
Accuracy of SVM: 0.5657254138266796
Accuracy of Random Forest: 0.5532821824381927
Accuracy of Decision Tree: 0.4390451832907076
Accuracy of GBM: 0.5268542199488491
Accuracy of Logistic Regression: 0.44757033248081
```

Methodology

- Data Preparation: Merge Questions, Answers, and Tags datasets.
- Preprocessing: Remove HTML tags, lowercase text, lemmatize words, and eliminate stop-words.
- Feature Engineering: Apply TF-IDF to extract features.
- Sentiment Analysis: Calculate sentiment polarity using TextBlob.
- Modeling: Train machine learning models and measure accuracy.



Load Data

Preprocess Data

Vectorize and Encode

Train-Test Split

Train Models

Evaluate Models

Print Accuracies

Conclusion

By combining sentiment analysis, NLP, and machine learning, we effectively classify tags in Stack Overflow data. Techniques like lemmatization, TF-IDF, and sentiment scoring enhanced feature quality, while SVM and Random Forest proved highly effective. Future work will explore advanced NLP and deep learning methods for improved accuracy and scalability.

Supervisor: Prof Dr. Usman Ghani

Muhammad Usman Asghar 2021-CS-46