Muhammed Akgöçmen

CS412

10 January 2025

Project Report: User Classification and Like Count Prediction

---

Introduction

This project aimed to classify users based on their social media activity and predict the like counts of posts using machine learning models. The data included Turkish text captions, user information, and post statistics such as captions, like counts, and categories. The analysis involved preprocessing the data, feature engineering, and testing different machine learning models to achieve optimal performance.

Data Overview

1. **Source of Data**:
    - User-post relationships and captions were analyzed for classification.
    - Like counts were utilized for regression tasks to predict post engagement.
2. **Challenges**:

- Data inconsistencies, including missing captions or like counts.
- Handling Turkish text, including specific stop words, hashtags, mentions, and emojis.

Data Preprocessing

1. **Text Preprocessing**:
    - Lowercased text while removing emojis, mentions (@username), and hashtags (#hashtag).
    - Removed URLs, special characters, and extra whitespace to ensure clean text input.
    - Utilized Turkish-specific stop words for improved feature extraction.
2. **Handling Missing Values**:
    - Dropped rows with missing captions or like counts during preprocessing.
    - Users with insufficient data were excluded from the analysis.

Modeling Approach

1. User Classification

- **Goal**: Classify users into predefined categories based on their post captions.

- **Models Tested**:

    - **Random Forest Classifier**: Achieved reasonable accuracy with feature tuning.
    - **XGBoost Classifier**: Slight improvement in performance but slower.
    - **BERT**: Abandoned due to high computational cost and marginal improvement over simpler models.

- **Best Model**:

    - The **Random Forest Classifier** was chosen for its balance of performance (accuracy: ~0.59) and computational efficiency.

- **Key Observations**:

    - Classification accuracy was limited (~0.59 validation accuracy).
    - Incorporating additional features, such as emoji counts and hashtag frequency, improved performance marginally.

2. Like Count Prediction

- **Goal**: Predict post like counts based on captions and user-level features.

- **Approach**:
    - Normalized like counts by dividing each post's like count by the user's average like count.
    - Trained a Ridge Regression model on normalized counts and un-normalized during prediction.
- **Best Model**:
    - **Ridge Regression** provided interpretable results with reasonable RMSE.
    - Log-transformation was initially explored but abandoned in favor of normalization for better interpretability.
- **Key Observations**:
    - Normalizing like counts improved prediction consistency across users with different engagement levels.
    - Features such as hashtags, mentions, and emojis could provide additional predictive power.

Findings and Insights

1. **User Classification**:

- Classification accuracy was constrained due to overlapping user behavior and limited dataset size.
- Incorporating domain-specific features (e.g., emoji usage) helped capture nuanced behavior patterns.

2. **Like Count Prediction**:

   - Normalizing like counts by user-specific averages allowed the model to generalize better.
   - The Ridge Regression model was computationally efficient and yielded stable results.

3. **Modeling Trade-offs**:

   - Simpler models (Random Forest and Ridge Regression) performed comparably to complex models (BERT and XGBoost) with significantly lower computational costs.
   - Feature engineering and domain knowledge were crucial for improving model performance.

Recommendations

1. **For Future Work**:

- Explore ensemble methods combining models (e.g., Random Forest and XGBoost) for better classification accuracy.
- Use pre-trained embeddings like FastText for capturing semantic information in Turkish text.
- Fine-tune models on a larger dataset to mitigate the impact of data scarcity.

2. **Practical Applications**:

- Use the classification model to segment users for targeted content strategies.
- Leverage like count predictions to optimize posting strategies for higher engagement.

Conclusion

This project demonstrated the effectiveness of traditional machine learning models (Random Forest and Ridge Regression) in tackling user classification and like count prediction tasks. While deep learning models like BERT were tested, their computational costs outweighed their benefits given the dataset size and performance gains. Future work could focus on integrating additional domain-specific features and leveraging larger datasets for enhanced accuracy.