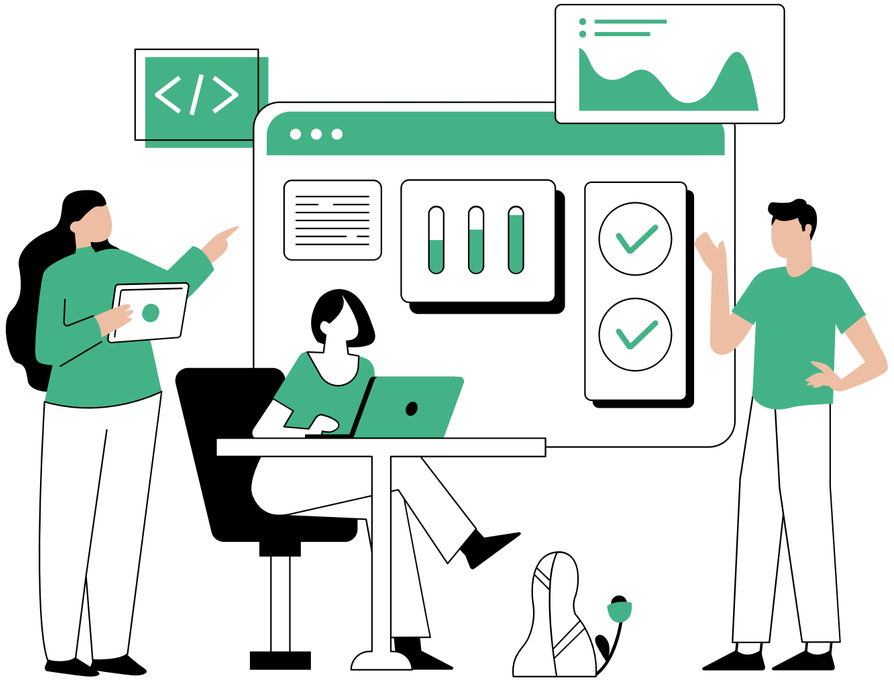
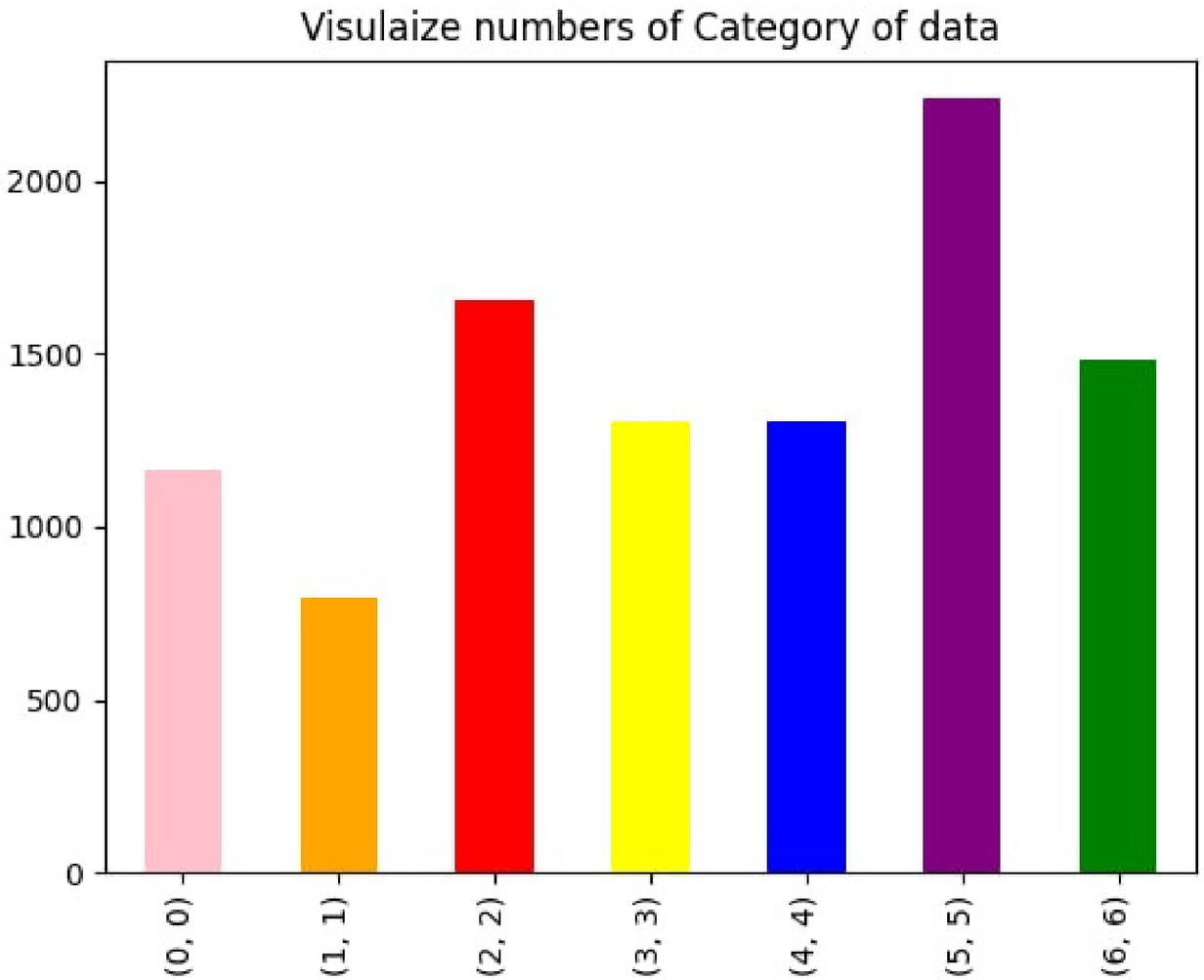
Classifying iLur News

Team: habenama



Task:

The task is to classify news articles by their topic.
The dataset contains around 12,428 texts from iLur.am, with 7 different topics: art, economy, sport, accidents, politics, society, and weather.



Our train dataset preprocessing includes:

- Converting all letters to lowercase
- Removing punctuations and numbers
- Removing stop words
- Stemming

Logistic regression

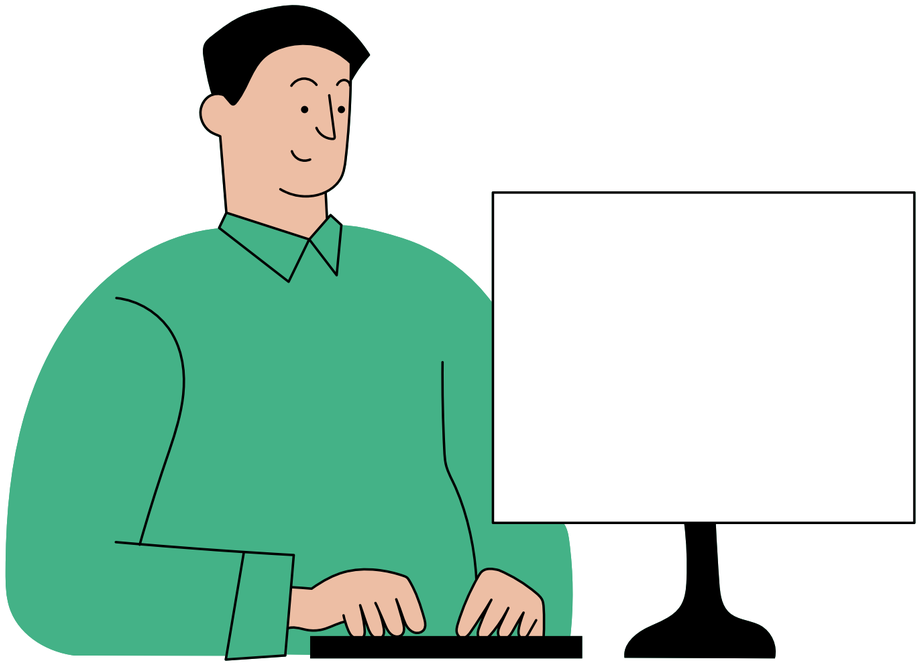
What are the advantages of Logistic Regression?

1. Logistic Regression is very easy to understand.
2. Logistic regression can easily be extended to multi-class classification
3. They are easier to implement, interpret, and very efficient to train.

What are the disadvantages of Logistic Regression?

1. Logistic Regression requires a large dataset and also sufficient training examples for all the categories it needs to identify.
2. It is quite sensitive to noise and overfitting.

Our best result with LR



For the start we preprocessed our dataset with removing stop words and punctuation. Then we vectorized our dataset with TF-IDF. After that we trained our model with parameters: solver(liblinear), C(8), multi_class(multinomial), penalty(l2). And the result was 0.91384.

Final result
Logistic Regression CV

For the start we preprocessed our dataset with converting all letters to lowercase, removing stop words and punctuation. Then we vectorized our dataset with TF-IDF and scaled with MaxAbsScaler. After that we trained our model with parameters: cv(10), max_iter(200), n_jobs(4).
And the result was 0.91534.

Thank you!

May, 2022