Clickbait Detection on Turkish News Articles

Mehmet Mert Tezcan

Outline

- Problem Definition
- Dataset Characterization
- Research Questions
- Methodology
- Expected Challenges

Problem Definition

- Clickbait titles are misleading and they are not informative.
- Clickbait titles are used to attract users to click on the news article.
- By predicting clickbait titles, we can prevent users from being misled.

Dataset Characterization

- 20,036 news article titles collected from different news websites
- 10,030 clickbait titles and 10,006 non-clickbait titles
- There are 4 columns named as "id", "clickbait" (label), "site" and "title". Only the title and label are useful, but I can extract some more features from title.
- Source: Kaggle

Research Questions

- How can we convert the clickbait titles into a meaningful representation for the machine learning models?
- How can we extract some features from the clickbait titles?
- Which machine learning model is the best for clickbait detection? Do we also need to use deep learning models?

Methodology

- Data Preprocessing
- EDA (Exploratory Data Analysis)
- Feature Extraction
- Tokenization, Vectorization
- Machine Learning Models (Logistic Regression, Naive Bayes, SVM, Random Forest, etc.) I can try to use all and compare the results.
- Parameter Tuning, to improve the performance of the models.
- Evaluation

Expected Challenges

- NLP is not the topic of the course
- Implementing it for Turkish language could be a challenge