UNIVERSITÀ
DI PAVIA

Cake classification

D. Ligari 518592[1]

[1] *Machine Learning course, University of Pavia, Department of Computer Engineering (Data Science), Pavia, Italy*
Github page: https://github.com/DavideLigari01/cake-classification
Contact: davide.ligari01@universitadipavia.it
Date: June 13, 2023

Abstract — This project focuses on building a browser capable of differentiating clickbait headlines from regular headlines using machine learning techniques. A dataset consisting of 32,000 headlines, equally divided into clickbait and non-clickbait classes, has been collected for training and evaluation. The dataset is provided in text files, with one headline per line. The project involves several tasks, including data analysis, designing and implementing a data pre-processing procedure, training and evaluating classification models, and analyzing the behavior of the trained models using data processing and visualization techniques. Two scenarios are considered: a generic scenario where all errors are equally important and a precision-oriented scenario that prioritizes minimizing false positives.

Keywords — Clickbait headers recognition • Naive bayes • Logistic regression • Bag of words • Stemming

CONTENTS

1. INTRODUCTION

In the era of online content consumption, clickbait headlines have become a prevalent means of attracting users' attention and enticing them to click on a link. These headlines are often deceptive, sensationalized, or misleading, failing to accurately represent the actual content being delivered. To address this issue, a software house has embarked on a project to develop a browser capable of distinguishing clickbait headlines from regular headlines. This project aims to design and implement a classifier using machine learning techniques that can accurately predict whether a given headline is a clickbait. The tasks involved in the project include data analysis, data pre-processing, model implementation, training, evaluation, and analyzing the behavior of the trained models using suitable data processing and visualization techniques. Additionally, two scenarios are considered: a generic scenario where all errors are equally important, and a precision-oriented scenario that aims to minimize false positives.

2. THE DATASET

The dataset used in this project consists of 32,000 headlines, which have been collected and labeled for training and evaluation purposes. The headlines are divided equally into two classes: 'clickbait' and 'non-clickbait'. Clickbait headlines are those that are intentionally crafted to capture attention, often using sensationalized or misleading language. On the other hand, non-clickbait headlines are expected to be more informative and accurately reflect the content they are associated with.
The dataset is further divided into three sets: a training set, a validation set, and a test set. The training set comprises 24,000 samples, which will be used to train the classification model. The validation and test set consist of 4,000 samples.
The data is stored in text files, with each headline occupying a single line.

3. FEATURES SELECTION

Considering that the dataset consists of headlines, a bag-of-words approach is deemed the most appropriate choice for feature selection. This approach considers each word as a feature, resulting in a vector representation of each sample, in which each element represents the number of occurrences of a given word in the headline.
Furthermore, several variations were applied, including the removal of stop words, which are the most common words lacking specific meaning. This step aims to eliminate noise and focus on more meaningful and informative words.
Additionally, stemming was performed to reduce words to

their basic form, disregarding variations due to tense or plural forms. This normalization process enhances the capture of word essence, independent of superficial grammatical differences.
Multiple models were trained for each of the variants, and their results were compared to identify the best combination of features. The objective was to determine the most effective feature selection approach that maximizes the classification model's performance.

4. MODELS USED

a. Naive Bayes

Naive Bayes is a popular and efficient classification algorithm based on Bayes' theorem. It assumes that features are independent given the class label. The algorithm calculates the probability of a sample belonging to each class and assigns the class with the highest probability. Naive Bayes aims to minimize misclassification by selecting the class label that maximizes the posterior probability.

size of the vocabulary

Features selection

b. Logistic regression

Logistic Regression is a widely used supervised machine learning algorithm for binary classification. Unlike linear regression, which predicts continuous values, logistic regression models the probability of an instance belonging to a particular class. It estimates the relationship between the input features and the log-odds of the target class using a logistic function.
In logistic regression, the model's parameters are learned by maximizing the likelihood function or minimizing the log loss (also known as cross-entropy loss) during training. The log loss measures the dissimilarity between the predicted probabilities and the actual class labels. By minimizing the log loss, logistic regression aims to accurately classify instances and maximize the model's predictive performance.

Choice of the learning rate

Choice of the vocabulary size

Features selection

Number of iterations

c. Comparison between the two models

5. ANALYSIS OF THE BEST MODEL

a. Confusion matrix

b. Most impactful words

6. DECLARATION

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.