UNIVERSITÀ DI PAVIA

Machine Learning Course

I'm Not Clickbaiting You... It's a Headline, I Swear!

A machine learning solution to clickbait

Andrea Alberti

Department of Computer Engineering - Data Science

University of Pavia, Italy

Email: andrea.alberti01@universitadipavia.it

GitHub: https://github.com/AndreaAlberti07/Clickbait-Detection-ML.git

June 16, 2023

Abstract

This report explores the implementation of classification algorithms for clickbait detection, with a focus on comparing the performance of the Multinomial Naïve Bayes Classifier and Logistic Regression models. The objective of the project is to evaluate the effectiveness of these models in identifying clickbait content using two different approaches: an accuracy-oriented approach and an approach targeting the minimization of False Positive Rate (FPR).

The experimentation of various combinations of vocabulary size and type, led to the selection of models that demonstrate impressive results. In the accuracy-oriented scenario, the models achieved a test accuracy of 97.12%. In the FPR-oriented scenario, a 0% FPR was achieved on the test set, while maintaining a good accuracy of 84%.

Additionally, this project provides valuable information about the composition of clickbait headlines. It identifies the most impactful words for the classification models, shedding light on the characteristics that make headlines challenging to classify accurately. The analysis of the worst errors further enhances the understanding of the limitations of the model.

Contents

1 Introduction

Clickbait headlines have inundated the digital landscape, flashing users with sensational promises and generating significant challenges for content consumers and platforms. To address this issue, classification algorithms have emerged as a solution for identifying clickbait content automatically. This report explores the implementation of these algorithms, comparing their effectiveness and performance. By understanding clickbait detection advancements, it is possible to tackle the issue of misleading headlines and improve the reliability of online information.

2 Goal

The objective of this project is to train and compare the performance of the *Multinomial Naïve Bayes Classifier* and *Logistic Regression* in clickbait detection. The project focuses on two scenarios: the first is an accuracy-oriented approach, aiming to maximize the classifier's overall accuracy. The second scenario is FPR-oriented, prioritizing the minimization of False Positive Rate (FPR) while maintaining a satisfactory level of accuracy.

3 Data

The data consists of a dataset of 32,000 headlines. The dataset is evenly divided into two classes: 'clickbait' and 'non-clickbait'. It comprises three subsets: training, validation, and test sets, containing 24,000, 4,000, and 4,000 samples, respectively. The data is stored in text files, with one headline for each line.

3.1 Data Pre-processing

To prepare the data for the algorithms it is necessary to convert the headlines into a numerical representation (features extraction). One common way tor this is to create the *Bag of Words* representation. This process involves two steps:

- **Build a vocabulary**: the vocabulary is a sub-list of all the words that appear in the training set.
- **Build the BoW**: each headline is represented as a vector of length equal to the vocabulary size. Each element of the vector is the number of times the corresponding word appears in the headline.

For this project two types of vocabulary were created: one with stopwords removed and one without, each one with different sizes. The punctuation was removed, but the numbers were kept.

4 Classification Models

The two used models follow different philosophies. The Naïve Bayes Classifier is a generative model, while Logistic Regression is a discriminative one.

4.1 Multinomial Naïve Bayes

As a generative model, it classifies data by first learning $P(Y = y)$ and $P(X = j|Y = y) = \pi_{y,j}$ independently for each class y. Specifically, it assumes a multinomial distribution for $P(X|Y = y)$, thereby it classifies according to the following rule:

$$\hat{y} = arg\max_{y} \sum_{j=0}^{n-1} x_j \log \pi_{y,j} + \log P(Y) \tag{1}$$

4.2 Logistic Regression

As a discriminative model, it classifies data by learning the boundary directly. Specifically, it learns the function 2 and estimates $P(Y = 1|X = \mathbf{x_i})$ as 3.

$$z_i = b_i + w_1 x_{i1} + w_2 x_{i2} + ... + w_n x_{in} \tag{2}$$

$$\hat{p_i} = \frac{1}{1 + e^{-z_i}} \tag{3}$$

5 Scenarios

The primary concern in certain use cases may not always be the overall accuracy of the model, but rather minimizing the number of false positives. For instance, in clickbait detection, misclassifying legitimate headlines as clickbait can lead to user frustration. Thus, it is important to prioritize reducing the number of false positives. Indeed accuracy remains important, therefore a reasonable tradeoff needs to be found. In this project, for each scenario, the two models were trained and evaluated using different sizes of the two vocabulary types (with and without stopwords).

5.1 Max Accuracy Oriented

The results of this scenario are summarized in Figure 1, where the model accuracies are compared changing vocabulary size and type.

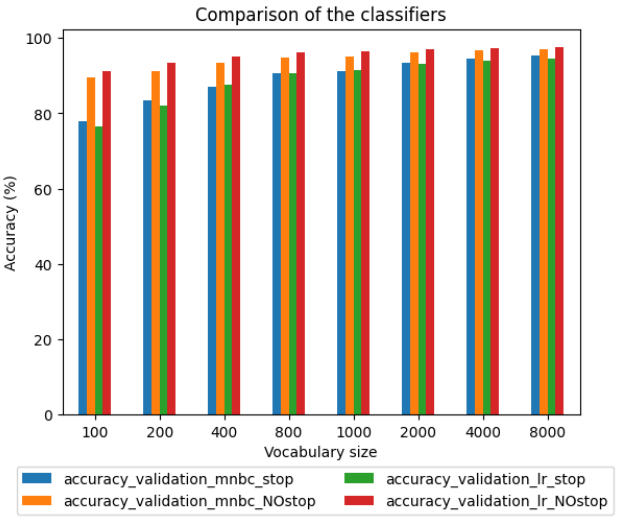

Figure 1: Accuracy for all models

A first outcome is about the **vocabulary type**. The models trained with the vocabulary without stopwords ('_stop') perform worse than the ones trained

with the vocabulary with stopwords ('_NOstop'). This indicates that these latter may be useful. Moreover, it is worth to underline that each headline is made up by only a few words, therefore removing stopwords may lead to a loss of information. A second results is about the **vocabulary size**. In general, there is a positive correlation between the size of the vocabulary and the validation accuracy. As the vocabulary size increases, the model's ability to capture a wider range of linguistic patterns and variations improves, as witnessed by Figure 2. However, it's important to note that this correlation tends to plateau after a certain dimension. Regarding the **model type**, both the Multinomial Naïve Bayes Classifier and Logistic Regression models demonstrate similar performance. However, the Logistic Regression model requires more time to train due to the optimization process and hyperparameters choice[1]. Consequently, the Naïve Bayes Classifier is preferred due to its faster training time, allowing for selecting the model with the largest vocabulary size. Finally, its efficient training time, coupled with the validation accuracy of 97% make the NBC with stopwords and vocabulary size of 8000 the best model for this scenario.

5.2 Min FPR Oriented

To find the model with the lowest FPR, the biases were varied in the range [-8, 8] for LR and [5, -5 ; -5, 5] for MNBC. Subsequently the algorithm 1 was appositely developed and applied.

Algorithm 1 Finding the Model with the Lowest FPR

```
for each Model do
    for each Vocabulary Type do
        for each Size in Vocabulary Sizes do
            for each b' in biases do
                Train Model → w, b
                w, b' → Make Inference on Validation
                Compute FPR
            end for
            Take smallest FPR, related b' and accs
        end for
    end for
end for
```

The results are presented in Figure 3 and Figure 4. Once again, it is evident that the models that do not remove stopwords (bottom chart in the figures) outperform the others, limiting our choice to these models. Both the MNBC and LR models exhibit impressive performance, achieving an FPR of 0.0005 (0.05%) while maintaining a high accuracy above 84%. While some instances show a 0% FPR, it is important to note that these results may vary with small changes in the test set, therefore a difference of 0.05% is not considered significant. Considering the aforementioned advantages of fast training and the favorable trade-off between accuracy and fpr, the **MNBC with a vocsize of 2000** is chosen as the best model. Note that by picking another bias[2], the model can be tailored to the specific case.

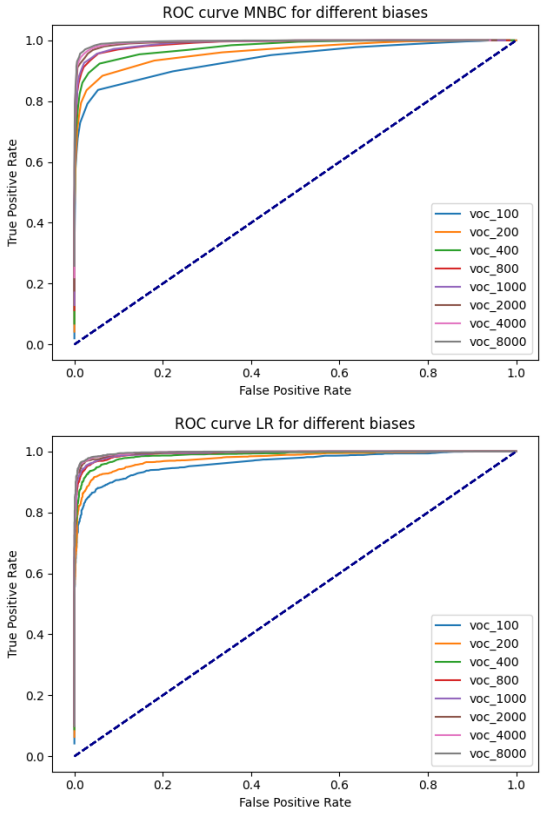

Figure 2: ROC curves for models keeping stopwords

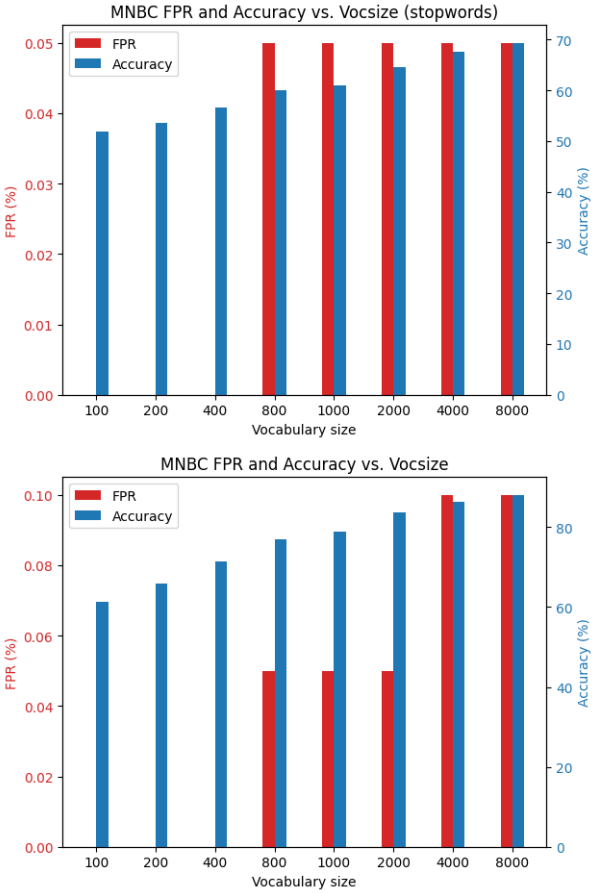

Figure 3: FPR and validation_accs for MNBC

[1]The details about the Logistic Regression setup are reported in section 7

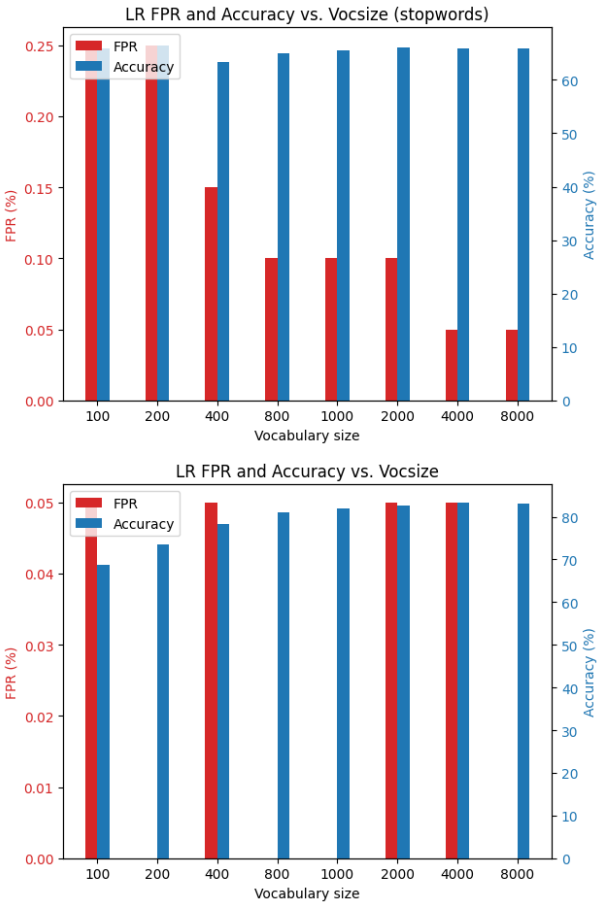

Figure 4: FPR and validation_accs for LR

The results of the best model are summarized in Table 1.

Vocsize	biases	FPR	TPR	Val_acc
2000	5,-5	0.0005	0.672	83.58%

Table 1: Best Model FPR scenario

6 Best Models Analysis

Based on the previous sections, the best model have been identified for each scenario are:

- **Max Accuracy Oriented**: MNBC with stopwords kept and vocabulary size of 8000
- **Min FPR Oriented**: MNBC with stopwords kept, vocabulary size of 2000 and biases = [5, -5]

In this section, a comprehensive analysis of the best models is presented, including an investigation into their worst errors and the identification of the most impactful words on the test set.

- **Most impactful words**: the most impactful words for a class are those that most influence the classification score in favour of that class. They are taken looking at the probability gap of each word between the two classes.
- **Worst errors**: these are the data misclassified with the higher confidence. In other words, are the headlines for which the gap between the classification scores of the two classes is the largest.

Finally, the models are evaluated on the test set, and the results are documented.

6.1 Max Accuracy Oriented

Most impactful words

The results are reported in Table 2. The words in the "Bait" column, with their negative deltas, have been determined to be strongly associated with clickbait headlines. These words may include terms commonly used in clickbait headlines, such as specific years ('2015'), attention-grabbing phrases ('hilarious'), or words that provoke curiosity or emotional engagement ('guess'). When these words appear in a headline, they tend to increase the likelihood of it being classified as clickbait. On the other hand, the words in the 'Nobait' column, with their positive deltas, are associated with non-clickbait headlines. These words may include terms related to news topics ('iraq', 'afghanistan'), informative language ('announces'), or general topics of interest ('nuclear'). When these words are present in a headline, they tend to decrease the likelihood of it being classified as clickbait. The model learned to assign weights and deltas to words based on their frequency and association with clickbait or non-clickbait labels in the training data. Through this process, the model identified these impactful words as important features that contribute significantly to the classification decision.

Bait	Delta	Nobait	Delta
2015	-5.70	kills	5.56
things	-5.69	iraq	5.36
these	-5.24	wins	4.88
tweets	-5.02	afghanistan	4.88
you	-5.01	leader	4.85
guess	-4.97	wikinews	4.68
hilarious	-4.76	announces	4.64
actually	-4.76	zealand	4.59
ve	-4.62	nuclear	4.58
instagram	-4.57	iraqi	4.55

Table 2: Most impactful words for the best ACC scenario

Worst errors

In Figure 5 are shown respectively the false positives and the false negatives on the test set, together with the headlines content and the model confidence. The misclassified headlines seems reasonable, since even a human could have some doubts about their classification. Despite the length of the headlines might appear as an influencing factor, as the false positives are shorter than the false negatives, it is not the case as depicted in Table 3. This analysis is useful to provide information about how, the headlines capable of deceiving the model, are composed.

[2]Further results available at this GitHub repository

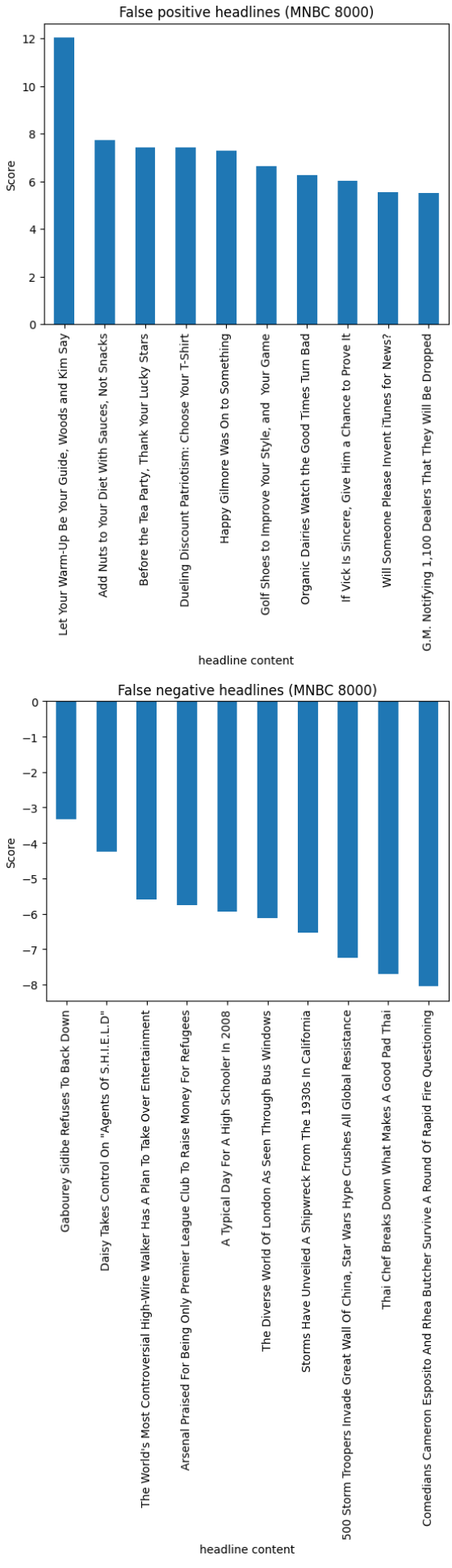

Figure 5: Worst errors for MNBC (8000)

Misclassified	All	Correctly Classified
8.65	9.11	9.12

Table 3: Avg length (n. words) of the headlines

Results on Test set

The results reported in Table 4 confirm the impressive performance of the model, achieving a test accuracy of 97.12% despite its simplicity.

Train Accuracy	Test Accuracy
97.88%	97.12%

Table 4: Test results for the best ACC scenario

6.2 Min FPR Oriented

Most impactful words

The most impactful words for this scenario are reported in Table 5. It is interesting to note that the words are the same as the ones for the best accuracy scenario, with slightly different deltas. This indicates the robustness of these features in identifying clickbait headlines. Specifically, it suggests that these words carry substantial information about the clickbait nature of the headlines and can serve as reliable indicators for classification purposes.

Bait	Delta	Nobait	Delta
2015	-5.54	kills	5.72
things	-5.53	iraq	5.52
these	-5.08	afghanistan	5.04
tweets	-4.86	wins	5.04
you	-4.84	leader	5.01
guess	-4.81	wikinews	4.84
hilarious	-4.60	announces	4.80
actually	-4.60	zealand	4.75
ve	-4.46	nuclear	4.74
instagram	-4.41	iraqi	4.71

Table 5: Most impactful words for the lowest FPR scenario

Worst errors

In the analysis of the worst errors, the false negatives are not reported since they are the same as those for the best accuracy scenario. Moreover, it is noticeable that the model did not produce any false positives on the test set. This outcome highlights the effectiveness of the model in minimizing the False Positive Rate (FPR). According to Table 6 the TPR is 68% which means that the model is able both to avoid false positives and recalling 7 out of 10 clickbait headlines.

It is worth mentioning that the relation between FPR and TPR is expressed by the ROC curve and can be tuned at any time by changing the bias, tailoring the model to the specific use case.

Results on Test set

The results reported in Table 6 confirm the excellent performance of the model, achieving a False Positive Rate of 0% on the test set.

Test Accuracy	FPR	TPR
84.00%	0.0%	68%

Table 6: Test results for the lowest FPR scenario

7 Details: Logistic Regression

In this section a brief overview about the Logistic Regression setup is provided.
All the results in the previous sections were obtained training the model using the full batch gradient descent algorithm, with a learning rate of 0.001 and 1000 iterations. No regularization was applied.
The training was limited to 1000 iterations since a convergence trend was observed in the loss function. This decision was made to strike a balance between ensuring convergence and allowing for the exploration of multiple scenarios within a reasonable training time frame.
The learning rate was selected from a range of options, including 0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, and 0.1. The choice was guided by observing the behavior of the loss function during training, specifically looking for a steep and consistent decreasing trend without oscillations.
Since the model did not exhibit overfitting, displaying a small discrepancy between the training and validation accuracy, no regularization techniques were employed.

For seek of completeness the LR with a vocabulary size of 4000 and keeping stopwords, was trained using the following specifications:

- **Learning rate**: 0.001
- **Iterations**: 20000
- **Regularization**: 0.2
- **Tolerance**: 0.0000005

Looking at the results depicted in Figure 6 and Figure 7, they are evident both the convergence of the model after few iterations and the small gap between the training and validation accuracy.

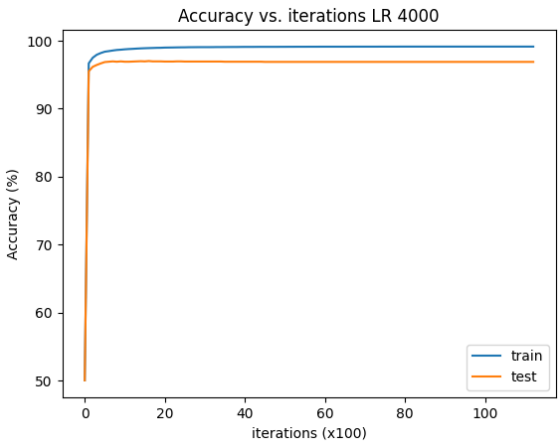

Figure 6: Accuracy for LR

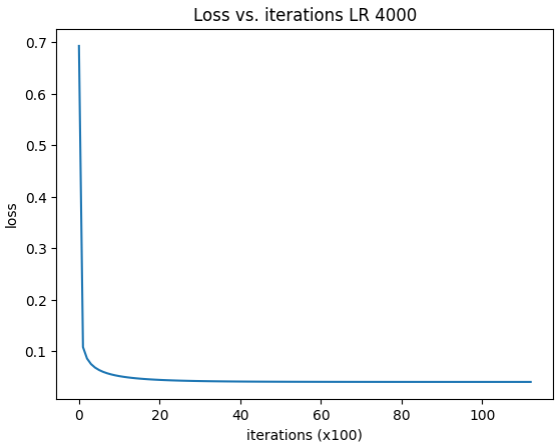

Figure 7: Loss function for LR

8 Conclusions and Future Work

Future Work

The inclusion of numbers in the vocabulary is important for the clickbait classification model as clickbait headlines often contain numeric values. However, in the list of impactful words, can be noticed that only one number appears. This is because the model relies solely on the frequency of a specific word in the clickbait and non-clickbait classes, without recognizing whether a word is specifically a number or not. Since headlines can include a wide range of different numbers, the model is unable to identify a specific number as impactful. Adding a feature to identify numbers in headlines could significantly enhance the already good model's performance.

Conclusions

In summary, the outcomes obtained in this project can be deemed satisfactory. Remarkable performance was observed in both scenarios for the tested models. Among them, the Multinomial Naïve Bayes Classifier emerged as the preferred choice due to its ability to deliver notable results with minimal training efforts. The composition of the vocabulary was identified as a crucial element in influencing the model's performance, with the inclusion of stopwords yielding superior outcomes. Additionally, the results underscored the significance of vocabulary size, indicating that larger vocabularies correlate with improved performance.
Lastly, the project offered information about the structure of headlines capable of fooling the model, and revealed that attention-grabbing and emotionally engaging words are commonly used for clickbait headlines.

The code and further results of this project can be freely accessed and reviewed on the GitHub repository at AndreaAlberti07/Clickbait-Detection-ML.git.

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.