Test task

Makar Charviakou

October 31, 2023

1 Introduction

In this task, I needed to analyze data related to marketing and simple publications and answer some questions:

1. What are the most common languages?
2. Which countries have the highest reach?
3. What sources provide the greatest return on marketing investment?
4. What happened when columns had negative sentiment values?

In order to answer all these questions, I preprocessed the data (filled NaNs, created some new features), and finally generated some plots.

2 My code and additional datasets

You can access all this on my GitHub page: GitHub.

3 Explanation of my code and thought flow in detail.

3.1 Data inspection

First of all, we always need to inspect our data: see distributions in numerical columns, check for NaN's, fill them with some values and find out if some columns are connected (correlated) with each other.

Note: My work is more complete and less effective, and if needed, it's possible to delete some preprocessing sections to save time.

1. NaN's Our data has a lot of NaN values, which in the future can cause different problems with the code, and it's easier and more pleasant to work with clear data. In Table 1, you can see the count of NaN values for each column.
2. Correlation between numeric columns in our data.

   I wanted to see if our metrics have anything connected with likes/shares/comments/days/weekdays. Unfortunately, we have only a strong correlation between likes/shares/comments (because of NaN's and also due to the fact that they all reflect some action in the publication).

   This correlation matrix has name: Figure 1.
3. Dealing with unnecessary columns

   First of all, "Language" and "Language.1" are the same columns, that's why we can delete one of them. Secondly, column "Unnamed: 0" is an index column. We don't need it.

4. Country column

   In one of our tasks, we need to take into account the country column. The problem is that we have mostly every country column values filled with NaNs. My solution is to take languages in which people accessed our website and replace the country column with them. It's not a precise solution, but I think one of the quickest and most intuitive.

5. Distributions in "shares," "likes," "comments," "host traffic" columns

   We also need to fill NaNs with some values here. These are numerical columns and they reflect some interactions of users with publications. I assume that it's not good to fill these NaNs with 0.The minimal reach of our publications is 74, and the median is 129,920, indicating that there must be some interaction. I decided to fill all NaNs with the medians of non-NaN values because they have high variance in our columns (Figure 2). In the end, we can sum up all the previously mentioned columns to get the number of interactions per post.

6. Filling NaNs in categorical columns

   The easiest and clearest way to deal with this problem is to simply fill out values with the most frequent values.

7. Filling NaNs in "lemmas" column

   On the way to an ideal dataset, we encounter a more complex problem: NaNs in the "lemmas" column. We can't simply fill these lemmas with some values because the "lemmas" column consists of words in the "content." The best way to deal with it is to use some tools like lemmatization and stemming, which are widely used in the NLP field. The final result of this preprocessing appears to be good.

8. Idea with future potential

   I also want to share my previous idea about the 'emoticons' column. I thought it's a good idea to make use of columns "emoticons," "emotion," "sentiment" columns and create a column that accumulates all these columns to give a more precise result.

9. In the end, I created a "sentiment mood" column which replaces the category "neutral" with 0, "positive" with 1, and "negative" with -1.

In the end of this section, we finally got an ideal-looking dataset without NaNs.

4 Answering Questions

In this part, I will answer questions asked in the introduction part.

4.1 What are the most common languages in this dataset?

Here (Figure 3), we can see that the most common languages are **Polish, second is English. Next are Slovene, German, and Hungarian languages**.

4.2 Which countries have the highest reach?

On Figure 4, we can see that countries with the highest reach are **Poland, Worldwide (can be any country), Germany, and Serbia**.

4.3 What sources provide the greatest return on marketing investment?

To determine the most effective sources, I considered the "AVE" (Advertising Value Equivalent) metric, which is highly correlated with return on investment "ROI" and also both "AVE" and "ROI" have same sense: measure effectiveness of marketing company or publication. Figure 5 displays the top sources by AVE, which are **Samagame, social networks (Facebook, Instagram, TikTok), Samagamer, and Upflix**.

4.4 What happened when columns had negative sentiment values?

Here, I created two plots. In the second plot, I tried to implement the created metric "mood". The result not good. My metric didn't capture one day when a problem occurred. I think this happened because emojis are very ambiguous and need to be interpreted in context. Figure 6. shows the distribution of good and bad interactions based on the "sentiment" column, and Figure 7 displays the same distribution based on the "mood" column. I identified specific days with negative values:

1. **Day 286 (2023/10/13): Most of the comments mentioned missing the second episode of "Loky". on Disney+.**
2. Day 295 (2023/10/22): Many comments discussed films, which might be coincidental.
3. **Day 296 (2023/10/23): Users reported problems with the application, likely due to software issues.**

5 Conclusion

In this test task, we conducted a comprehensive analysis of marketing and publication data, focusing on language prevalence, reach by country, effective marketing sources, and the impact of negative sentiment values.Also tried to use new features and improved quality of data.

6 Plots

Table 1: Missing Values in Each Column

Column	Missing Values
created_date	0
url	0
title	847
language	0
author_name	4574
content	0
host	0
host_traffic	6
emotion	345
sentiment	176
intent	306
lemmas	221
shares_count	3302
likes_count	3006
hashtags	3021
comments_count	3034
emoticons	4681
country	4719
ave	0
reach	0

Table 2: Final look at our data

Does the column have NaN's?	**Answer**
created_date	False
url	False
title	False
language	False
author_name	False
content	False
host	False
host_traffic	False
emotion	False
sentiment	False
intent	False
lemmas	False
emoticons	False
country	False
ave	False
reach	False
day_of_year	False
num_of_interactions	False
sentiment_mood	False
mood	False

Figure 1: Correlations for numeric columns

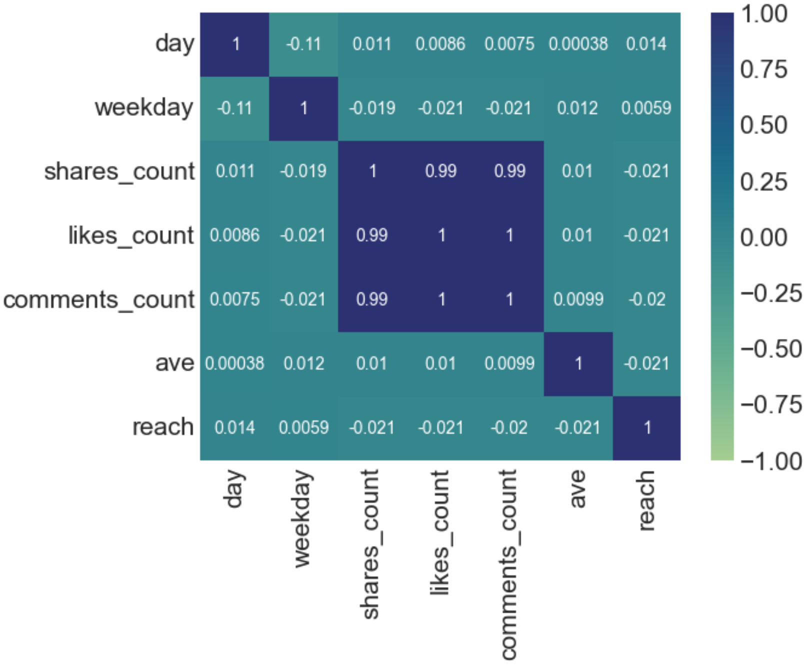


Figure 2: Distribution of interaction columns

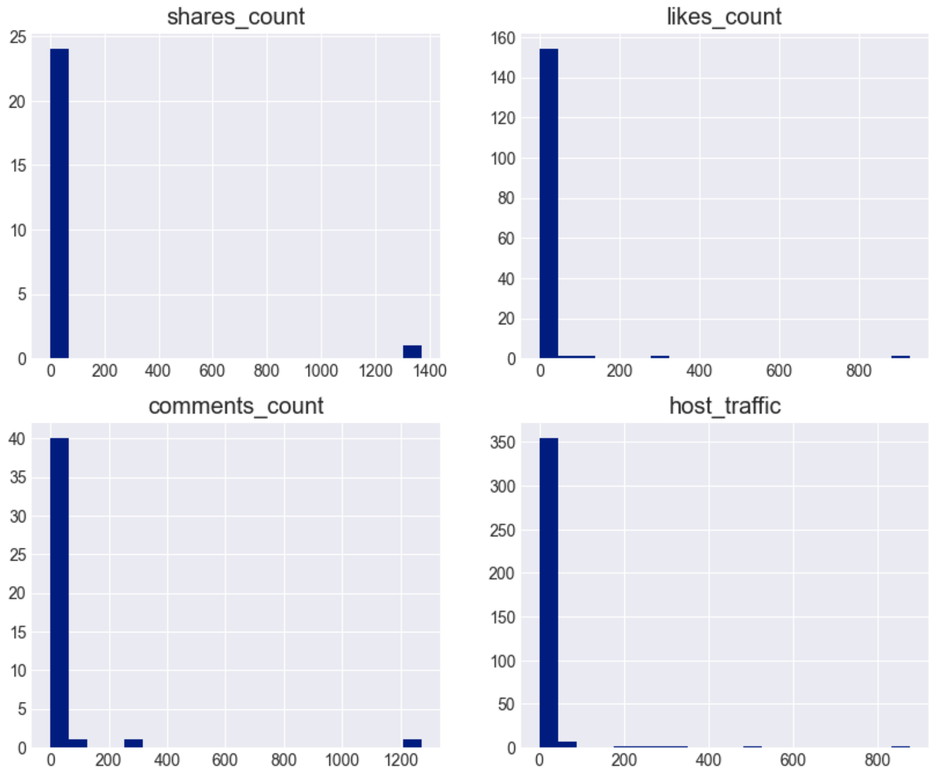


Figure 3: Bar plot for languages

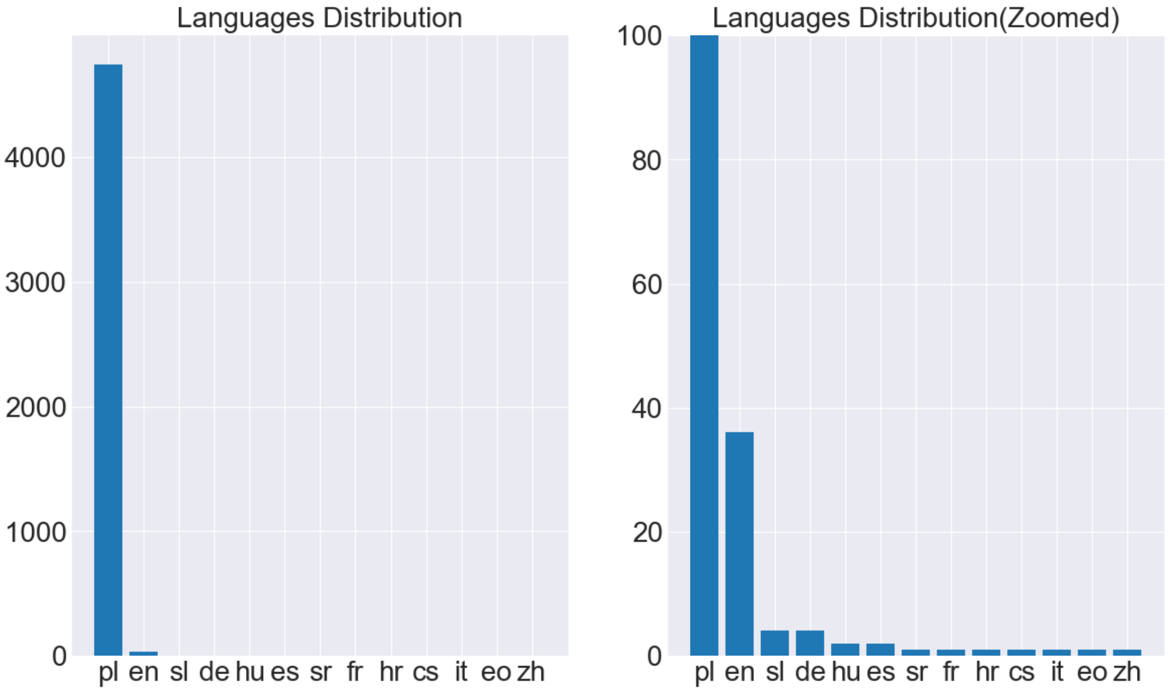


Figure 4: Bar plot for reach in countries

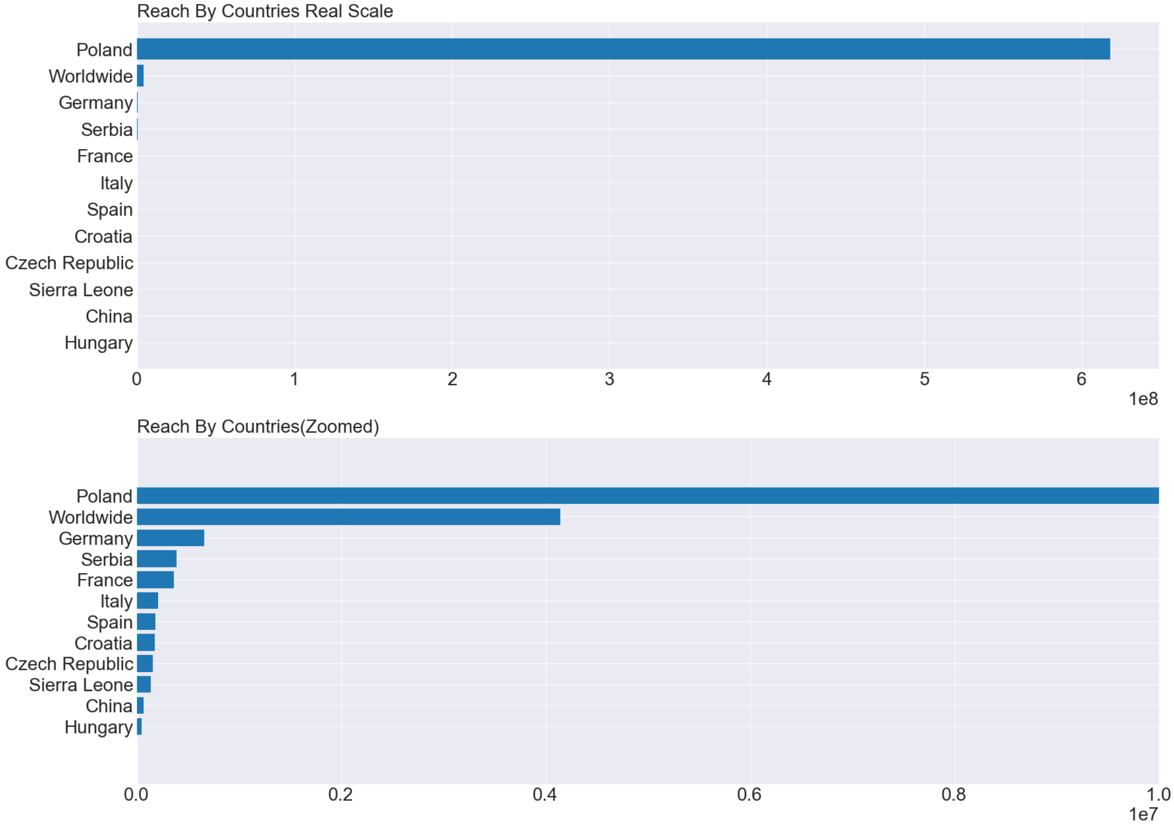


Figure 5: Top sources by AVE (Top 1-20 and Top 20-40)

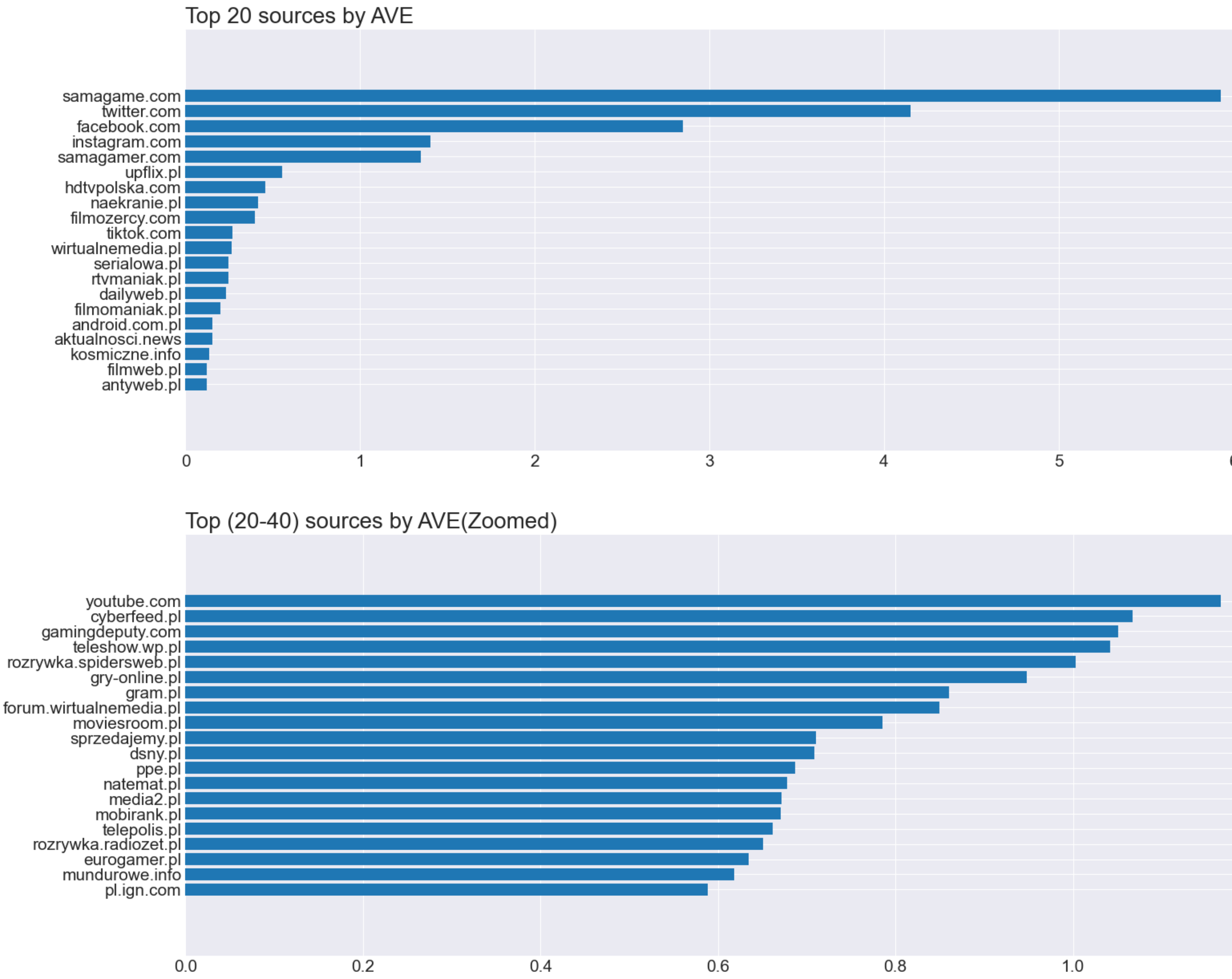


Figure 6: Bar of good or bad interactions by ”sentiment” column

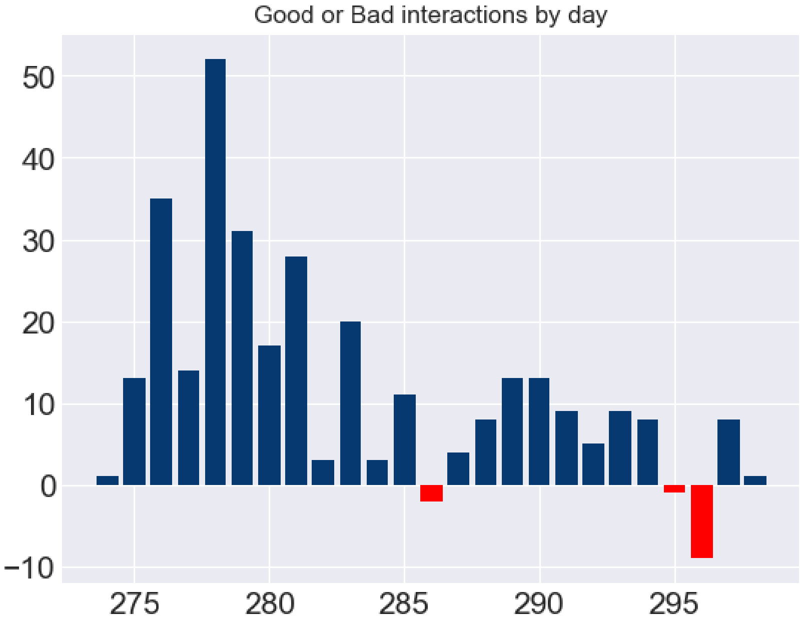
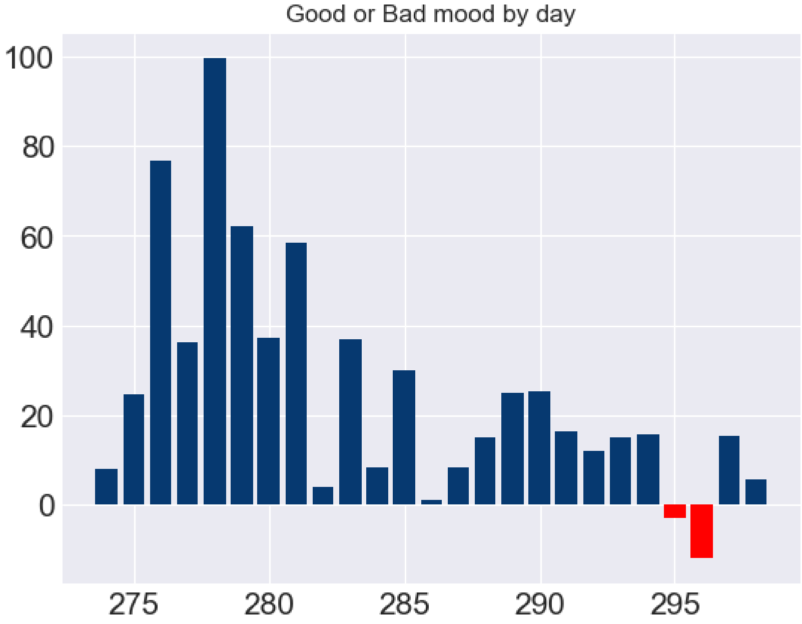


Figure 7: Bar of good or bad interactions by ”mood” column


Good or Bad mood by day