Part A – NLP

This project involves classifying text data, which consists of 21 distinct classes. For each entry, four pieces of information are provided: the URL path, the title of the article, the content of the article, and its corresponding class. For classification reason during CSV reading every class was replaced with a specific class id.

Two solutions have been implemented: the first solution yields a coarse result, while the second provides a finer result. Both solutions utilize the same NLP algorithm, namely the BELT pretrained multilingual classifier. Scikit-learn libraries are employed to extract and process the data. The evaluation metrics include validation and training accuracy, as well as loss, to determine the efficiency of the models. At test set f1-score was calculated for every class. In this way there is a clearer preview of True Positive, True Negative, False Positive and False Negative data.

At low latency solution only the title of the article was used for tokenization reasons. At high accuracy solution the content of the article was used. (Potentially “Content” and “Article” data could be used for tokenization but this solution is not implemented at this research for time saving)

The research commenced with an initial exploration of the model to identify and specify its parameters. These parameters, crucial to the model's performance, are substantially influenced by the dataset employed. Among the parameters under scrutiny, two were selected for focused investigation: the learning rate and the batch size.

The learning rate plays a pivotal role in determining the trajectory of the optimization algorithm. It dictates the size of steps taken during the optimization process. Optimal learning rate selection is crucial, as it directly impacts whether the algorithm converges towards the global minimum or gets trapped in a local minimum. A small learning rate may lead the algorithm to converge slowly and potentially stall in local minima, while a large learning rate may cause it to overshoot the global minimum. Therefore, meticulous adjustment of the learning rate is essential for efficient convergence and optimal model performance.

In addition to the learning rate, the batch size is another parameter that significantly influences the performance of the model. The batch size determines the number of samples utilized in each iteration of the training process. A larger batch size often leads to more stable and efficient convergence due to leveraging more information per iteration. However, excessively large batch sizes may lead to computational inefficiencies or hinder the model's ability to generalize well to unseen data. Thus, careful consideration and experimentation with batch size are imperative to strike a balance between computational efficiency and model performance.

In summary, the investigation of model parameters, particularly the learning rate and batch size, is essential for optimizing model performance. Through meticulous adjustment and experimentation with these parameters, researchers can ensure efficient convergence towards the global minimum while maintaining computational efficiency and generalization capabilities.

Model Investigation

Tokenization on "Content"

Model1

At model 1 the learning rate of the model is 5e-5, epochs number is 10 and batch size is 8.

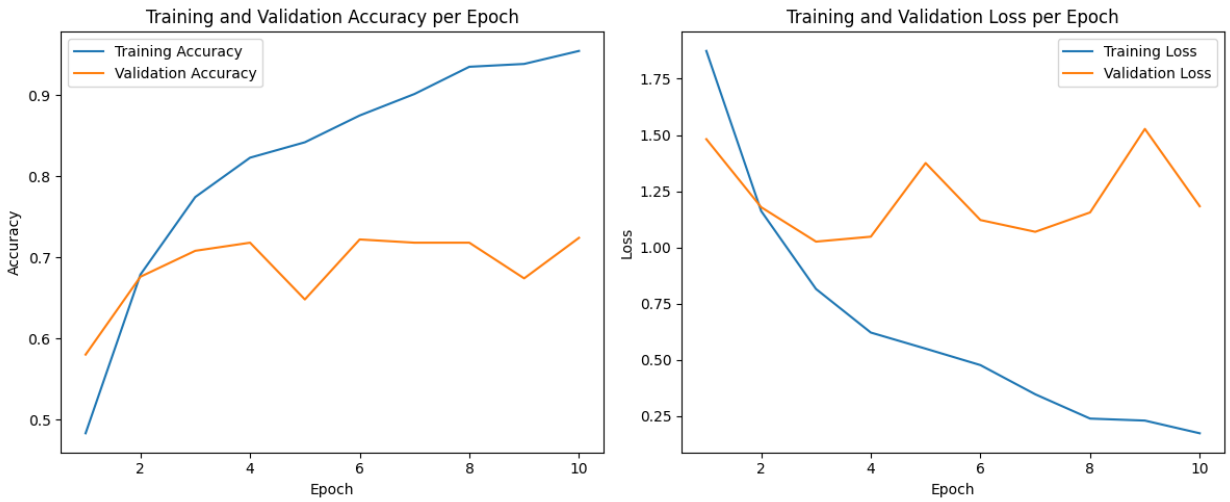

Figure 1.Model1: a) Training and validation accuracy curves b) Training and validation loss curves.

The analysis from Figure 1 reveals that the learning rate employed is adequate for the dataset. However, upon examining the accuracy and loss curves of the validation set, it's apparent that they lack smoothness and exhibit frequent fluctuations. Furthermore, it's noteworthy that the optimal model performance is achieved at the 6th epoch. Beyond this epoch, there is a discernible tendency for the model to overfit the data.

Model 2

At mode2 the learning rate of the model is 5e-6, epochs number is 10 and batch size is 8.

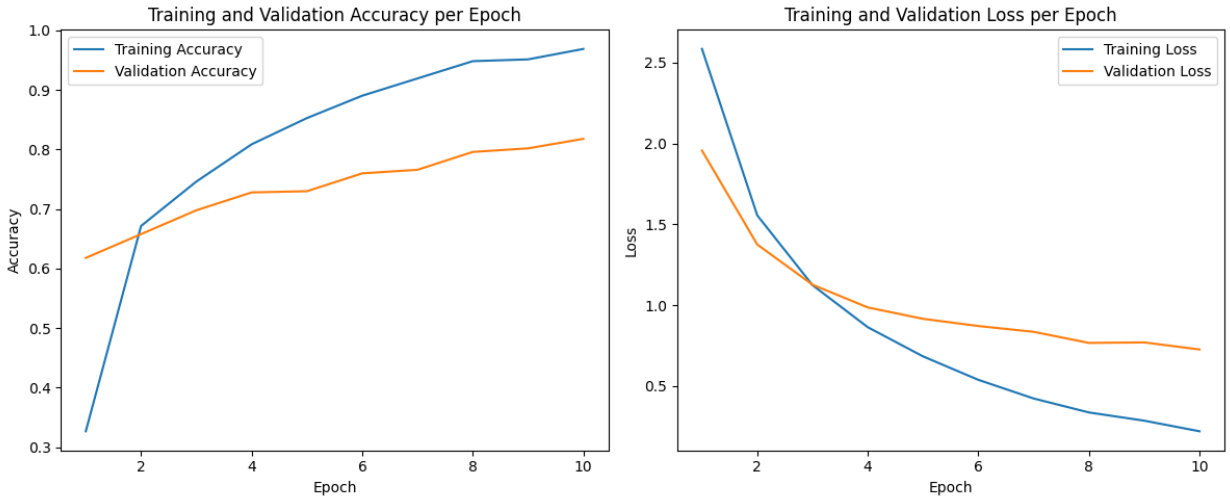

Figure 2. Model2: a) Training and validation accuracy curve b) Training and validation loss curves.

Based on the insights drawn from Figure 1, an adjustment is made to the learning rate of Model 2, reducing it to 5e-6. This alteration yields promising results. The curves exhibit enhanced smoothness, suggesting an improved learning trajectory. Notably, the model demonstrates potential for further learning beyond the 10th epoch, as indicated by the absence of plateaus in both accuracy and loss curves. Moreover, the validation accuracy approaches approximately 80%, signifying notable progress.

Model 3

At mode3 the learning rate of the model is 5e-7, epochs number is 10 and batch size is 8

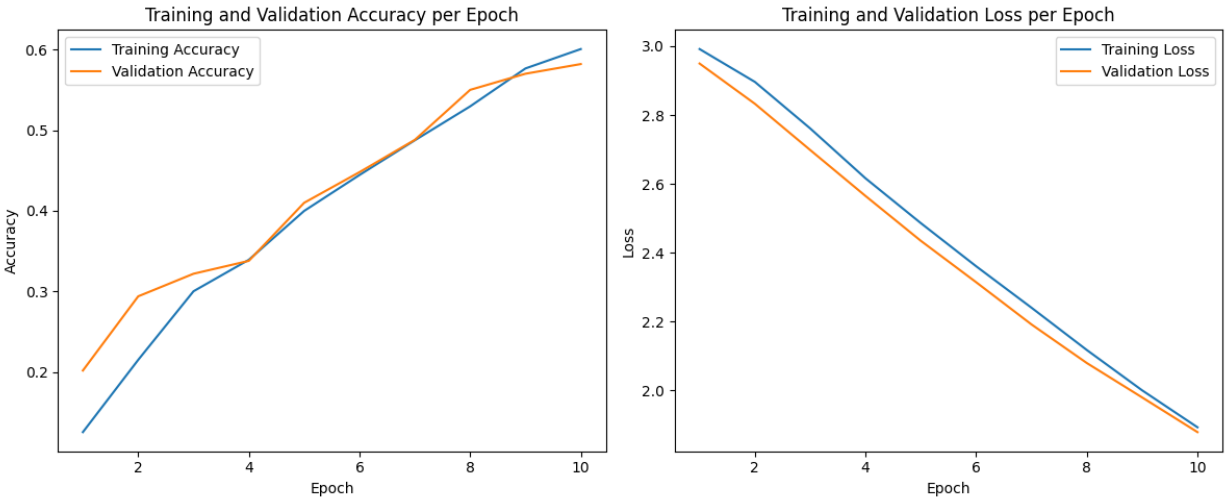

Figure 3. Model3 a) Training and validation accuracy curves b) Training and validation loss curves.

In Model 3, the learning rate was further reduced to 5e-7. However, it's noteworthy that this adjustment led to significantly longer execution times compared to Models 1 and 2. While the training and validation curves exhibit a similar alignment, indicating robust generalization to unseen validation data, the performance of Model 3 falls short compared to the favorable execution time and accuracy achieved by Model 2. Consequently, Model 3 is deemed less favorable in light of these considerations.

Model 4

At mode4 the learning rate of the model is 5e-6, epochs number is 10 and batch size is 16

Having determined that a suitable learning rate for the model is 5e-6, the investigation shifts to assess the impact of batch size on the model's performance.

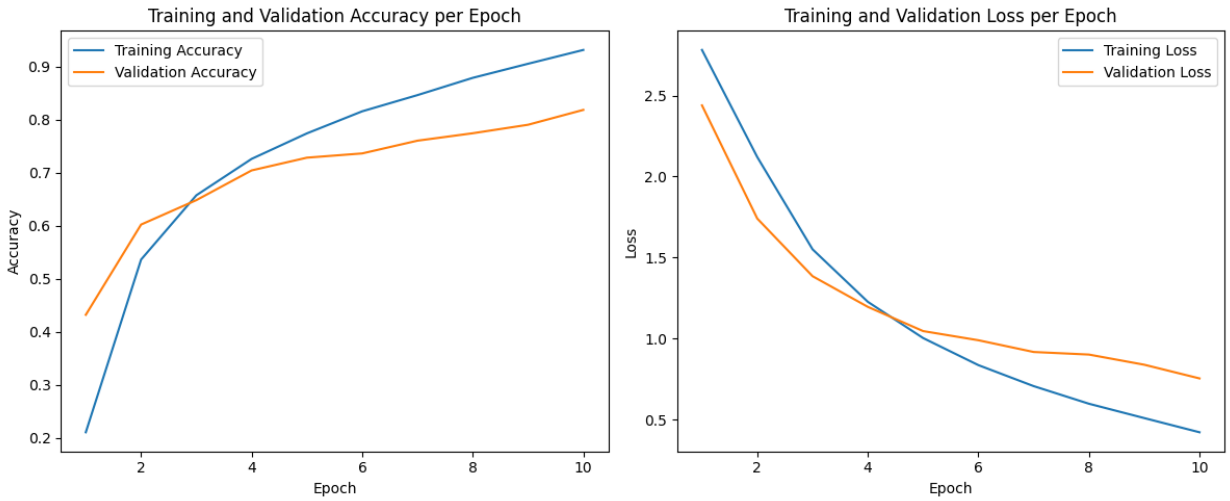

Figure 4. Model4: a) Training and validation accuracy curve b) Training and validation loss curves.

Comparing Model 4 to Model 2, it becomes apparent that the increase in batch size has a positive impact on the model's performance. Notably, this adjustment has led to a reduction in the disparity between the training and validation curves, evident not only in accuracy but also in loss curves.

Model 5

At mode4 the learning rate of the model is 5e-6, epochs number is 10 and batch size is 32.

Execution time increases dramatically, so for complex reason this model is rejected.

Conclusion

The learning rate of the model is 5e-6, epochs number is 10 and batch size is 32.

Tokenization on "Title"

For the low-complexity model, "tokenization with title" was selected based on parameters identified as efficient from the investigation. The model's learning rate is set at 5e-6, with 10 epochs and a batch size of 16.

The loss curves of the validation set show a decrease, accompanied by an increase in the accuracy curve. However, the differences observed are not substantial when compared to Model 2. Nonetheless, this model offers significantly lower latency compared to Model 2, resulting in reduced execution time requirements.

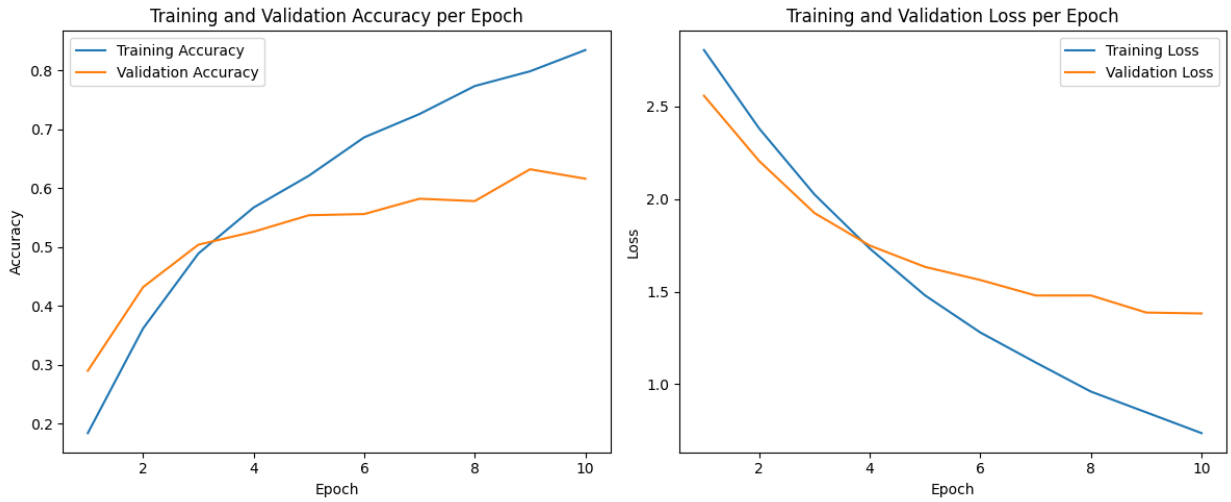

Figure 5 . Model5: a) Training and validation accuracy curve b) Training and validation loss curves.

Apply models on test set.

Table 1 F1- Score at test set

	F1 Score	
Label	Tokenization on "Content"	Tokenization on "Title"
Astrology-0	87.3%	74.3%
Attractions-1	72.3%	25.8%
Automotive-2	64.8%	77.5%
Beauty-3	51.7%	32%
Business&Finance-4	65.5%	39%
Culture-5	47.6%	31.8%
Education-6	92.5%	76.5%
Family&Relationships-7	57.6%	39%
Food&Drink-8	79%	55.7%
Healthy Living-9	48.4%	40.8%
Home&Garden-10	70%	0%
Politics-11	81%	54%
Pop Culture-12	71%	5.5%
Religion&Spirituality-13	78%	59.6%
Science-14	83%	56.3%
Sensitive Topics-15	40%	35.3%
Sports-16	84.7%	47%
Style&Fashion-17	52.5%	16.2%
Tech&Computing-18	89.6%	72.4%
Travel-19	55.4%	21%
Viral Articles-20	0%	0%

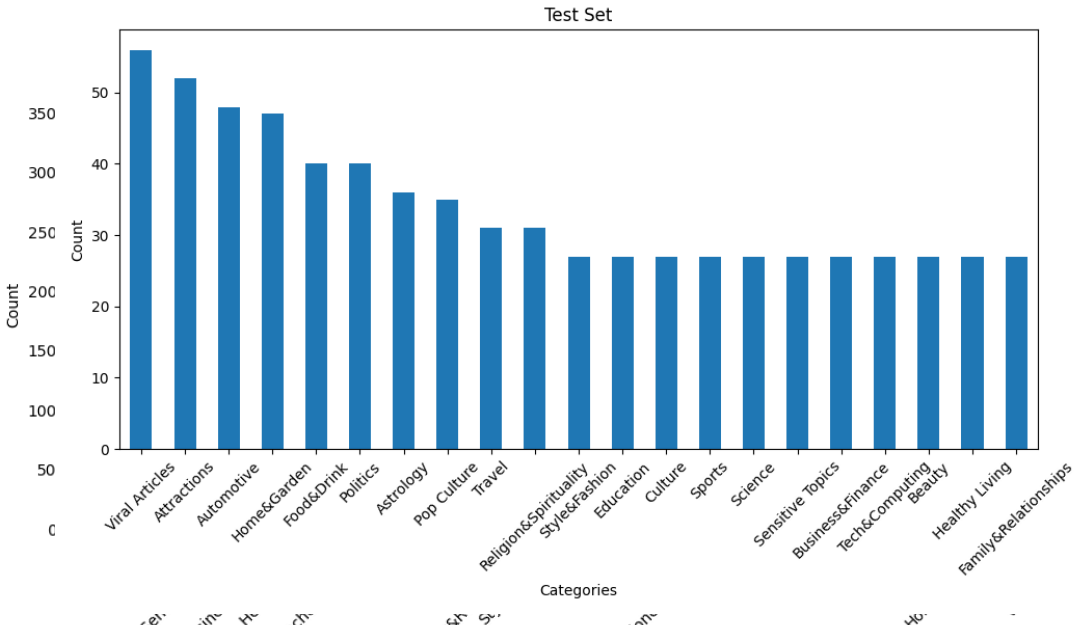

Figure 7. Class frequency on test set.

Figure 6. Class frequency on development set.

Remarks on dataset

Some classes exhibit low F1 scores, suggesting that our model struggles to accurately identify them. Therefore, it's essential to scrutinize our dataset for potential issues. Examination of the development set unveils significant class imbalance, where 11 out of 21 classes contain around 50 instances, while others boast as many as 360 data points. To mitigate model errors and bolster classification accuracy, a comprehensive analysis of the dataset is crucial. Addressing this class imbalance is imperative for achieving robust classification results.

Minimizing the errors of predicting a sensitive topic

If there is a requirement to minimize the errors of predicting a sensitive topic article as non-sensitive to zero, it introduces a significant constraint that needs to be addressed in the modeling approach. This requirement implies that false negatives (predicting a sensitive topic as non-sensitive) should be minimized as much as possible, potentially at the expense of other evaluation metrics like precision or overall accuracy. Here are some changes that may need to be made to the modeling approach:

- Adjusting Class Weights: In addition to handling class imbalances, you may need to adjust the class weights further to give more emphasis to sensitive topic articles during model training. By assigning a higher weight to the sensitive class, the model will be penalized more for misclassifying sensitive articles as non-sensitive.
- Threshold Adjustment: You might need to adjust the decision threshold for classification to prioritize sensitivity over specificity. By lowering the threshold, you increase the sensitivity of the model, making it more likely to classify articles as sensitive, even if it means accepting more false positives.
- Cost-Sensitive Learning: This approach involves explicitly incorporating the costs associated with misclassifications into the model training process. In this case, the cost of misclassifying a sensitive topic article as non-sensitive would be set to a very high value, effectively minimizing the occurrence of such errors.
- Ensemble Methods: Ensemble methods can be particularly useful in this scenario. By combining multiple models trained with different emphasis on sensitivity, you can create an ensemble that collectively minimizes the errors of predicting sensitive topics as non-sensitive while maintaining overall performance.
- Fine-Tuning Hyperparameters: Fine-tuning hyperparameters such as regularization parameters or learning rates may be necessary to optimize the model's performance under the constraint of minimizing errors for sensitive topics.
- Data Augmentation: Augmenting the training data with additional examples of sensitive topic articles can help improve the model's ability to recognize and classify such articles correctly.
- Custom Loss Functions: Designing custom loss functions that explicitly penalize misclassifications of sensitive articles as non-sensitive can be effective in addressing this requirement.
- Post-processing Techniques: Post-processing techniques such as calibrating model probabilities or applying additional rules to filter out false negatives can be applied to further reduce the errors of predicting sensitive topics as non-sensitive.

Overall, the modeling approach would need to be tailored to prioritize sensitivity while balancing other evaluation metrics and constraints, such as class imbalances and overall accuracy. This may involve a combination of adjusting model parameters, data preprocessing, and post-processing techniques to meet the specific requirements of minimizing errors for sensitive topic articles.

Part B – Data Analysis

Analyze the provided CSV file containing a sample of real-time bidding data to uncover valuable insights. Start by cleaning the dataset to ensure accuracy and reliability. Then, explore correlations between different variables. Ultimately, draw conclusions from the findings to strategize the company's next advertising steps effectively.

Data Cleaning

- Unique Values

  First of all developers should check if every column of the dataset has got unique values.

day	1
device	3
domain	231
utm_campaign	102
utm_content	33
utm_medium	57
utm_source	66
utm_term	27
country	116
oeid	13646
auction_start	14743
auction_end	14685
auction_ttl	4569
auction_id	16633
ad_unit_code	1285
bidder	17
bidder_cpm	1397
bidder_start	15953
bidder_end	13652
bidder_ttl	3166
bidder_is_after_timeout	2
bidder_media_type	4
bidder_width	11
bidder_height	13
bidder_source	1
bidder_status	3
bidder_deal_id	10
consent	2
advertisers	126
subdomain	20
is_impression	2
refreshed	2
os	7

Columns with unique values are not precious for current research.

- Check null data in the dataset.

Expressed as a percentage, indicate the proportion of missing values within the dataset.

day	0.000
device	0.000
domain	0.000
utm_campaign	97.045
utm_content	99.620
utm_medium	95.095
utm_source	95.190
utm_term	99.225
country	0.005
oeid	3.515
auction_start	0.000
auction_end	0.000
auction_ttl	0.000
auction_id	0.000
ad_unit_code	0.000
bidder	0.000
bidder_cpm	0.000
bidder_start	0.000
bidder_end	0.000
bidder_ttl	0.000
bidder_is_after_timeout	0.000
bidder_media_type	0.000
bidder_width	0.000
bidder_height	0.000
bidder_source	0.000
bidder_status	0.000
bidder_deal_id	0.000
consent	0.000
advertisers	0.000
subdomain	95.720
is_impression	0.000
refreshed	0.000
os	0.005

Result

Remove columns with high percentage of null data and column that all values are equal

- Hight percentage of nul: 'utm_campaign', 'utm_content', 'utm_medium', 'utm_source', 'utm_term','subdomain'
- All data are the same: 'day' 'bidder_source'

- ‘action_start’ and ‘action_end’ are refearing to the same date (2023-12-14)
- Data that refer to specific actions id are not taken into account because they do not provide extra information. Such data are oeid and auction_id.

Data Processing

- Countries Variance

Country is Greece :15048
Country is other: 4951
Data from Greece is about 75% out off all reported data.

Top 30 countries that are not “GR”

SI	1605
BG	633
CY	324
DE	288
US	263
EG	251
GB	192
RO	164
CA	94
NL	76
BE	59
IT	55
SE	52
FR	48
ES	46
CH	44
RS	44
AU	41
TR	41
AT	31
PL	31
GE	29
HR	25
DK	25
AL	24
UA	23
NO	22
AE	21
IE	21
ZA	20

It's noteworthy that approximately two-thirds of the events occur in Greece (GR). Following Greece, the next most popular country is Slovenia (SL).

- The top 5 most popular domains for the 10 most frequently visited countries.

The following tables present the countries alongside the total visits per domain.

Table 2. Countries and total visits per domain.

GR	SI	BG	CY	DE
sdna.gr 10.21%	bolha.com 29.66%	zajenata.bg 60.19%	argiro.gr 9.88%	argiro.gr 12.15%
argiro.gr 7.58%	metropolitan.si 27.41%	pik.bg 31.75%	greek-movies.com 7.10%	sdna.gr 11.46%
iefimerida.gr 5.40%	svet24.si 24.36%	168chasa.bg 5.85%	youfly.com 4.63%	greek-movies.com 5.90%
enikos.gr 3.49%	vecer.com 11.96%	sdna.gr 0.63%	enikos.gr 4.01%	gazzetta.gr 4.86%
zappit.gr 3.42%	nogomania.com 3.24%	zappit.gr 0.32%	bovary.gr 3.70%	marinetraffic.com 4.86%

US	EG	GB	RO	CA
alison.com 26.24%	masrawy.com 58.96%	whoscored.com 17.71%	b365.ro 84.76%	alison.com 25.53%
marinetraffic.com 25.10%	yallakora.com 33.86%	marinetraffic.com 14.58%	marinetraffic.com 6.10%	marinetraffic.com 22.34%
whoscored.com 11.79%	elconsolto.com 3.98%	sdna.gr 14.06%	whoscored.com 2.44%	whoscored.com 17.02%
argiro.gr 8.37%	marinetraffic.com 1.99%	alison.com 8.85%	alison.com 1.83%	greek-movies.com 7.45%
iefimerida.gr 3.80%	alison.com 1.20%	greek-movies.com 4.69%	sdna.gr 1.83%	sdna.gr 3.19%

Common domains among multiple countries:
zappit.gr: GR, BG
enikos.gr: GR, CY
argiro.gr: GR, CY, DE, US
iefimerida.gr: GR, US
sdna.gr: GR, BG, DE, GB, RO, CA
greek-movies.com: CY, DE, GB, CA
marinetraffic.com: DE, US, EG, GB, RO, CA
whoscored.com: US, GB, RO, CA
alison.com: US, EG, GB, RO, CA

- Device usage at 5 most popular countries.

The table below illustrates the devices used by individuals from the five most popular countries and their respective presence.

Table 3. Devices usage in the 5 most popular countries.

GR	SI	BG	CY	DE
mobile: 85.00%	mobile: 79.94%	mobile: 85.78%	mobile: 87.96%	mobile: 76.39%
Desktop: 14.28%	desktop: 20.00	desktop: 14.06%	desktop: 9.88%	desktop: 19.10%
tablet: 0.72%	tablet: 0.06%	tablet: 0.16%	tablet: 2.16%	tablet: 4.51%

- OS system usage.

Table 4. OS usage.

OS System	Usage (%)
Android	71.05%
Windows	13.68%
iOS	11.46%
OSX	2.45%
iPadOS	0.71%
Linux	0.48%
ChromeOS	0.17%

- Check the 10 most popular domain per device.

The plots below display the most popular domain per device.

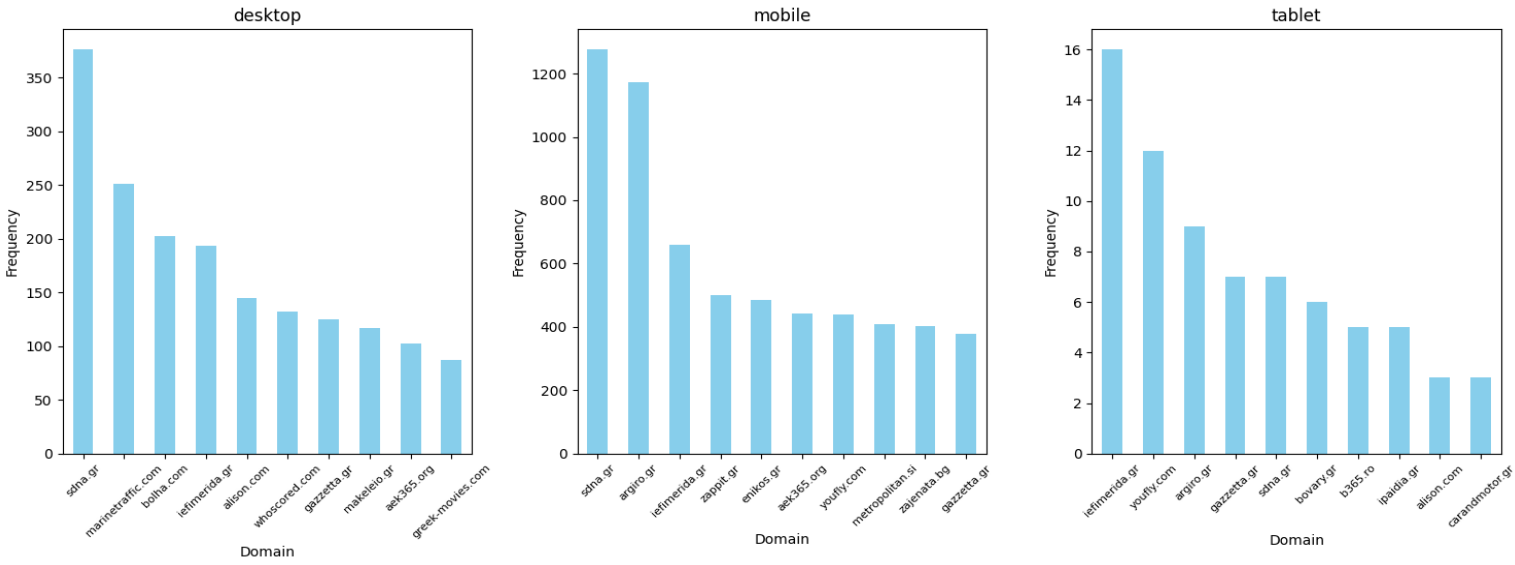

Figure 8. Domain frequency per device.

Common domains across all devices:
sdna.gr, gazzetta.gr, iefimerida.gr

- Highlighting Domains with the Longest Durations: Showcase the specific duration data for each domain in relation to the country where it is predominantly utilized.

Table 5. Domains with longest duration.

aek365.org	argiro.gr	iefimerida.gr	ieidiseis.gr	marinetraffic.com
GR 97.21%	GR 99.22%	GR 95.98%	GR 99.62%	US 99.89%
DE 0.46%	DE 0.24%	FI 1.08%	CY 0.26%	GB 0.03%
GB 0.46%	US 0.20%	US 0.63%	IE 0.06%	GR 0.02%

pentapostagma.gr	reader.gr	sdna.gr	zarpanews.gr
GR 98.02%	SE 93.45%	GR 93.55%	GR 99.93%
CY 1.17%	GR 6.03%	NL 1.49%	CY 0.07%
DE 0.73%	IT 0.17%	DE 1.47%	

- Top 3 advertisers.

Table 6. Top 3 advertisers.

Advertiser name	Frequency at data
advertisers0	98.89%
advertisers31	0.06%
advertisers26	0.06%

- Relation between advertiser and users' content.

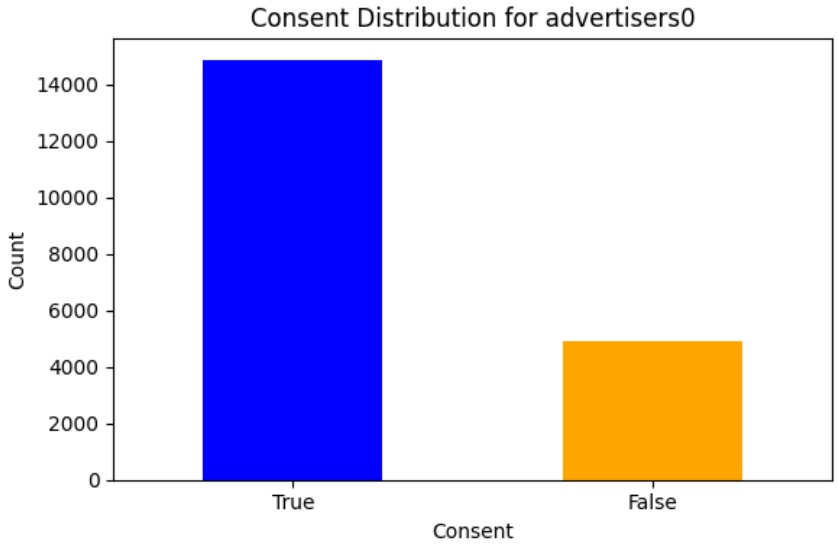

Figure 9. Users content (True /False) for advertiser0

- Advertiser0: 10 domains with longest duration.

Table 7. Domain's duration for advertiser0.

advertisers0	
Total duration for advertiser: 95597881.0	
Domains	Duration per domain
argiro.gr	48.78%
metropolitan.si	6.24%
aek365.org	3.29%
sdna.gr	3.17%
reader.gr	2.90%
iefimerida.gr	1.55%
ieidiseis.gr	1.53%
zarpanews.gr	1.40%
pentapostagma.gr	1.29%
marinetraffic.com	1.17%

- Total duration for each bidder_media_type .

Table 8. Duration for each bidder_media_type.

Bidder_media_type	Duration	Dominancy of bidder
banner	52.29%	59.27%
banner-video	35.08%	27.30%
video	12.63%	13.43%
NA	0.01%	0.01%

- Bidder media type, site and duration.

Table 9. Bidder media type, site and duration.

Bidder_media_type	Site	Duration percentage
banner	argiro.gr	47.13%
	reader.gr	5.39%
	sdna.gr	4.59%
	aek365.org	4.04%
	ieidiseis.gr	2.72%
banner-video	argiro.gr	65.36%
	pentapostagma.gr	3.09%
	marinetraffic.com	2.33%
	youfly.com	1.68%
	gossip-tv.gr	1.44%
video	metropolitan.si	44.72%
	argiro.gr	8.27%
	aek365.org	7.64%
	iefimerida.gr	4.86%
	sdna.gr	3.11%
NA	sdna.gr	100.00%

- Bidder with longest duration.

Table 10. Bidder duration.

Bidder	Duration (%)
bidder9	28.13%
bidder1	16.03%
Bidder3	15.19%
Bidder0	9.38%
Bidder4	5.19%
bidder10	5.18%
bidder7	4.02%
bidder8	3.62%
bidder6	3.52%
bidder5	3.49%
bidder2	2.20%
bidder12	2.17%
bidder11	1.64%

Conclusions

From the given dataset, several conclusions can be drawn. Firstly, approximately 75% of visitors to the domains are from Greece, where a diverse range of domains are accessed. Notably, sdna.gr

and argiro.gr are among the domains with high visitation rates, attracting both Greek and foreign visitors. Conversely, marinetraffic.com, alison, and whoscored.com primarily draw foreign visitors.

Furthermore, analysis of device usage statistics in the most popular countries reveals that visitors predominantly utilize their mobile phones to access the sites, accounting for around 80% of usage. Approximately 5% use desktops, while only 2% prefer tablets. Additionally, there's a significant preference for Android software, constituting about 71%, compared to Windows at 14%.

Examining specific domain preferences by device type, desktop users primarily visit sdna.com and marinetraffic.com, while mobile phone users frequent sdna.com and argiro.gr. Tablet users, on the other hand, primarily visit iefimerida.gr.

In terms of duration spent on the domains, prominent sites include aek365.org, argiro.gr, iefimerida.gr, marinetraffic.com, pentapostagma, readwer.gr, sdna.com, and zarpanews. Regarding advertisers, "advertiser0" dominates with approximately 99% of the data. Most users provide their consent for advertising. Notably, argiro.gr and metropolitan.gr boast the longest user durations for "advertiser0".

Moreover, the banner emerges as the dominant bidding type, accounting for 60% of activity. Among bidder types, argiro.gr demonstrates one of the lengthiest durations, with bidder ID "bidder9" standing out for its extended duration.

Top 5 bidders

Bidder name	Popularity at data
bidder1	13.05%
Bidder0	10.66%
Bidder5	9.20%
Bidder8	8.48%
Bidder4	8.43%

Bidder media type