Alterism:
Analyzing image description usage in mastodon.social

**Final project presentation by
Tommaso Marmo and Cristal Rivera**

Agenda

01

Project Domain

02

Contextual
Information

03

Project Scope &
Objectives

04

Methods

05

Data Analysis

06

Conclusions &
Future
Development

01

Project Domain

Alt Text and Accessibility in Social Networks

Important Definitions

Alt-text: Written descriptions for images, critical for accessibility.

- Helps blind or visually impaired users.
- Supports Universal Design principles.
- Adds context and meaning to images.

Clients: Interfaces for accessing social media platforms.

- Design impacts how users add alt-text to posts.
- This project focuses on trends and patterns, not specific clients.

Fediverse: A group of decentralized social networks connected by the ActivityPub standard.

Mastodon: A social network software using ActivityPub, widely used in the Fediverse.

Mastodon.social: A popular Fediverse instance analyzed in this project.

Did you know that only

0.1% of images in Twitter

contain alt text descriptions?

Leaving a majority of content inaccessible to visually impaired users

Source: ."It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible.

02

Contextual Information

Why?

Contextual Information

Why Fediverse?

Data Accessibility: Policies of major social networks make accurate data analysis difficult and costly.

Rising Popularity: Decentralized networks are gaining traction for being more ethical and human-centric.

Why Clients?

Interfaces that allow users to access social media platforms

Use of alt text: Clients play a significant role in shaping the user experience, including how users add alt text to images.

Why Accessibility?

Often overlooked issue, particularly on social media.

Intersection of data science and accessibility, aiming to shed light on this topic.

Highlight how **visually disabled users are often excluded** due to the lack of image descriptions

As of March 2023, **Mastodon had over 10 million registered users**. Its growth surged significantly following Elon Musk's takeover of Twitter, although its **decentralized nature makes user tracking more challenging** compared to centralized platforms.

Andre Louis
@FreakyFwoof@universeodon.com

Right. There's been a lot of talk about #AltText and it's place here on the fediverse of late.
Whether you're blind, deaf or sighted, I'd be curious to not only have you fill out the relevant choice, but feel free to respond to this with more detail if you like.
Do you make use of, benefit from or even find alt-text useful, whether for images or audio?

3%   I'm deaf/blind and I do benefit from alt-text

0%   I'm deaf/blind and I don't benefit from alt-text

79%   I'm sighted and I do benefit from alt-text

18%   I'm sighted and I don't benefit from alt-text

Tanya Harrison
@WirlOfWirls@mastodon.nz

Just wanted to thank all in this instance for adding image descriptions. To a totally blind person like me it means a lot to have a picture in my mind of your images, especially the animal pics.

474   677   💬 25 · December 3, 2022 at 9:51:00 AM

@stefan I'm no visually impaired, but sometimes I need a little help to fully understand social nuances in images. I've found ALTs useful once in a while.

03

Project Scope & Objective

Main Objective

**Analyze alt text usage in Mastodon.social
in relation to the clients used to publish posts on the platform.**

Scope

1. **Global use of alt text usage** within the dataset.
2. Pinpointing the **most popular clients being used**.
3. **Variation of alt text usage** across different clients.
4. Observing possible **trends or patterns linking client** usage and alt text usage.
5. **Assessing general inclination** to add image descriptions.
6. **Reflect** on how image descriptions **contribute to a more inclusive and welcoming online environment.**

04

Methods

Methods

Exploratory Data Analysis (EDA)

- Descriptive Statistics
- Visualizations: Histograms and Boxplots:

Correlation Analysis

- Relationships between variables were explored using both Pearson and Spearman correlation coefficients

Feature Engineering

- Data Preprocessing
- Logarithmic scale allows visualization of the broad range of post counts.

05

Data Analysis

Data Analysis

01   **Data Exploration**

02   **Global Analysis**

03   **Analyzing By Client**

04   **Exploring Alt Text features of the top clients**

1. Data Exploration: Structure & Dictionary

	client	posts	atxt_yes	atxt_yes_pct	atxt_no	atxt_no_pct
0	Web	8265	1438	17.4	6827	82.5
1	dlvr.it	5806	1	0.0	5805	100.0
2	Mastodon for Android	1891	270	14.3	1621	85.6
4	AboveMaidstoneBot	1339	0	0.0	1339	100.0
5	Mastodon for iOS	1051	227	21.6	824	78.3
...	...	...	...	...	...	...
261	socialbot	1	1	100.0	0	0.0
262	PhonocasterMusicShare	1	0	0.0	1	100.0
263	openvibe	1	0	0.0	1	100.0
264	iflaapp	1	0	0.0	1	100.0
265	Today's Dérive app task	1	1	100.0	0	0.0

variable	meaning
client	Name of the client the posts come from.
posts	Total number of posts containing images for that client.
atxt_yes	Number of posts containing alt text in *all* images.
atxt_yes_pct	The percentage of atxt_yes in relation to posts.
atxt_some	Number of posts containing alt text in *some* images.
atxt_some_pct	The percentage of atxt_some in relation to posts.
atxt_no	Number of posts containing *no* alt text in *any* images.
atxt_no_pct	The percentage of atxt_no in relation to posts.

```
Number of rows:                266
Number of columns:             8
Data size (rows×columns):      2128
```

1. Data Exploration: Description

	Data Type	Missing values?	count	mean	std	min	25%	50%	75%	max
client	object	False	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
posts	int64	False	266.0	112.240602	643.625189	1.0	2.0	8.0	39.75	8272.000000
atxt_yes	int64	False	266.0	21.473684	105.282555	0.0	0.0	0.0	4.00	1438.000000
atxt_yes_pct	float64	False	266.0	36.290411	45.079671	0.0	0.0	0.0	100.00	100.000000
atxt_some	int64	False	266.0	0.093985	0.622382	0.0	0.0	0.0	0.00	7.000000
atxt_some_pct	float64	False	266.0	0.060461	0.527118	0.0	0.0	0.0	0.00	7.142857
atxt_no	int64	False	266.0	90.672932	568.477040	0.0	0.0	2.0	20.00	6827.000000
atxt_no_pct	float64	False	266.0	63.649128	45.113843	0.0	0.0	100.0	100.00	100.000000

1. Data Exploration: Cleansing

Sort by posts

Sorting the data **based on the amount of posts,** showing the clients with **higher post at the top.**

Remove 'Unknown'

Third most used client is labelled as "unknown". There are 265 clients.

Validating %

- **Minimize precision error**: round all percentage-based columns to a single decimal.
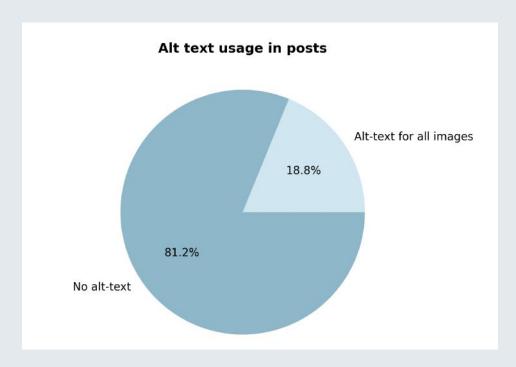- **Dataset are accurate and suitable for analysis** without requiring the creation of additional columns.

2. Global Analysis

Key Findings:

- **81.2% o**f posts **lack image descriptions**, while only **18.8% include them.**
- This aligns with Gleason et al. (2019), showing a lack of incentives for users to add alt text.
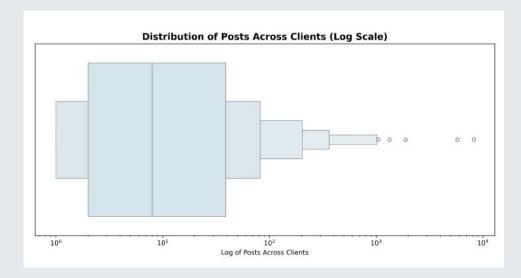
Comparison:

- Mastodon.social users use alt text far more often than Twitter users (**0.1% on Twitter vs. 18.8% on Mastodon**).

Alt text usage in posts

3. Analyzing by Client

Key Insights:

- **Spread data**: Most clients have low post counts, but a **few outliers post exceptionally high amounts**.
- The distribution **is skewed**: a small number of clients dominate total post activity.
- **High variability & Outliers:** Logarithmic scale shows wide differences in post counts.

Distribution of Posts Across Clients (Log Scale)

Log of Posts Across Clients

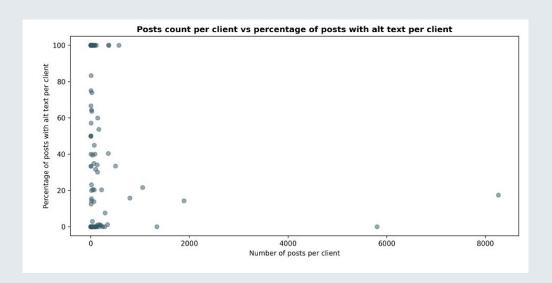|	
count	265.000000
mean	107.184906
std	639.324886
min	1.000000
25%	2.000000
50%	8.000000
75%	39.000000
max	8265.000000

3. Analyzing by Client

Key Observations:
- **No clear linear relationship** between the number of posts and alt text usage.
- **Most data points are around low post** counts and **varied alt text %**.
- Outliers show **clients with high post counts but lower percentages** of alt text usage.

Correlation Analysis
- Pearson measures the linear relationship between two variables.
- A value close to 0 indicates no linear correlation.
- Spearman measures monotonic relationships, value close to 0 suggests extremely weak positive monotonic relationship.

Posts count per client vs percentage of posts with alt text per client

```
Pearson correlation coefficient: -0.07
Spearman correlation coefficient: 0.02
```

3. Analyzing by Client: Why Grouping by Client popularity?

01

Clients with Few Posts

1. Limited data from clients with only a few posts **may not give reliable insights.**
2. **Noise or skewed results** due to random or inconsistent behavior.
3. **0.8% of the posts comes from the 43%** least popular clients.

02

Top Clients' Impact:

1. Clients with higher activity levels li**kely influence the overall trend in alt text usage**.
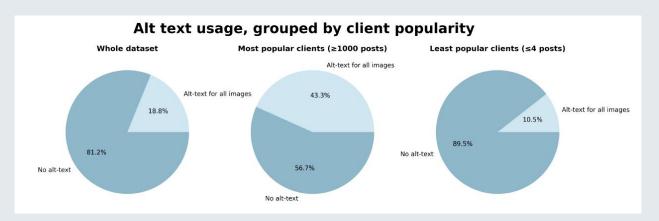2. **65% of posts come from the top 2%** of the clients.

03

Actionable Insights

1. Insights might be more valuable and can lead to **broader improvements in alt text accessibility** across platforms.

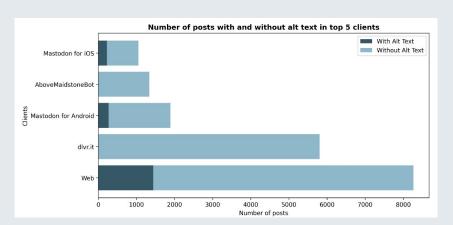# 3. Analyzing by Client: Grouping by Client popularity

Alt text usage, grouped by client popularity

Whole dataset

Alt-text for all images
18.8%

No alt-text
81.2%

Most popular clients (≥1000 posts)

Alt-text for all images
43.3%

No alt-text
56.7%

Least popular clients (≤4 posts)

Alt-text for all images
10.5%

No alt-text
89.5%

Overall Trend: Alt text usage is consistently lower than non-usage across all data.

Popular vs. Least Popular Clients

1. Most Popular Clients: Show lower alt text usage compared to other groups.
2. Least Popular Clients: Have higher alt text usage, but their impact is minimal as they contribute less than 1% of all posts.
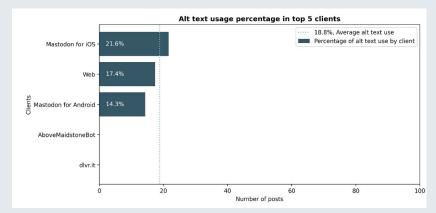
The most active clients contribute heavily to the dataset but lag in alt text usage, influencing overall accessibility trends.

3. Analyzing by Client: Grouping by Client popularity

Number of posts with and without alt text in top 5 clients

Alt text usage percentage in top 5 clients

Key Insights:

- While these **clients have the highest number of posts,** this **does not mean** they are more likely to **use alt text**.
- Suggests the **need for further investigation into how each client** encourages or supports alt text usage, impacting content accessibility.

Top Performer: *Mastodon for iOS* has the highest percentage of posts with image descriptions and is the only client above the dataset average.

Other Clients: *Web and Mastodon for Android* contribute a smaller but still relevant percentage of posts with alt text.
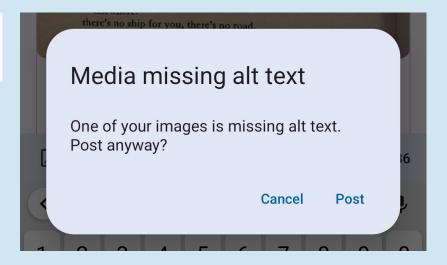
Lowest Performers: *AboveMaidstoneBot and dlvr.it* show no posts with image descriptions.

4. Exploring Alt Text features of the top clients

☑ Show confirmation dialog before sending toots lacking media descriptions

☑ Show confirmation dialog before boosting toots lacking media descriptions

ALT Add alt text reminders

there's no ship for you, there's no road.

Media missing alt text

One of your images is missing alt text. Post anyway?

Cancel   Post

06

Conclusions & Future Development

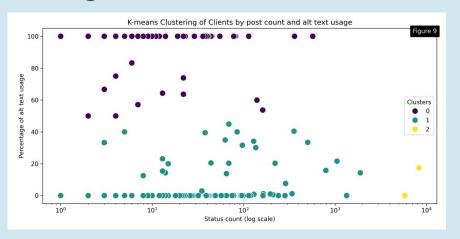# Evaluation Analytical Methods: Why Linear Regression, Logistic Regression, and KNN were not considered

Method	Considerations
Linear Regression	Revealed a very weak relationship (Pearson: -0.06, Spearman: 0.03), indicating a lack of linear association. Linear regression was unsuitable due to the risk of overfitting or underfitting in nonlinear data.
Logistic Regression	This method is for binary or categorical target variables. Most of the data (90%) were continuous metrics, like alt text percentage and post count. The goal was analysis, not binary prediction.
KNN	Requires a well-defined target variable for classification or regression. Creating a variable wasn't feasible due to insufficient significant features.
Alternative Approach	Explored grouping clients with similar behaviors using unsupervised clustering (Future Development).

Future Development: K Means Clustering

Objective: To identify grouping patterns based on status counts and the percentage of alt text usage No predefined labels or assumptions were needed to start.

Key Findings

- Cluster 2: Web and dlvr.it clients.Characterized by high post volumes but low alt text usage.
- Cluster 0: Mastodon for iOS, Mastodon for Android, and Mastodon.bot. Showed different patterns from Cluster 2.
- Silhouette score: 0.85
- Confirms meaningful clusters, offering a foundation for future projects and analyses.

K-means Clustering of Clients by post count and alt text usage

Figure 9

	client	status_count	descriptions_all_count	descriptions_all_percent
0	Web	8272	1438	17.383946
1	dlvr.it	5806	1	0.017224

6. Conclusions

Analysis Results

1. Many clients with very few posts.
2. The majority of posts with images does not have an image description, in all cases.

Challenges and Future Dataset Improvements

1. Updating recent data.
2. Gather information from other servers beyond mastodon.social.
3. Add more features.

Limited Alt Text Usage Despite Ethical Sensitivity:

1. Fediverse is perceived for more sensitive to ethical issues, analysis shows that image descriptions remain underutilized.
2. Gap in accessibility efforts that requires more attention.

Future Directions:

1. Real-world data often needs extensive cleaning and restructuring.
2. K-means clustering,quick start to set a foundation for future analysis and master's related topics.

DEMO :)

Thank you!