Zurich Like New: Analyzing Open Urban Multimodal Data

Marcel Granero-Moya
EPFL
Switzerland
marcel.graneromoya@gmail.com

Thanh-Trung Phan
Idiap Research Institute
Switzerland
tphan@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute and EPFL
Switzerland
gatica@idiap.ch

ABSTRACT

Citizen-driven platforms for enhancing local public services have been adopted in several countries like the UK and Switzerland. Local governments use data collected from these platforms to solve reported issues. Data can also be used by governments for data-driven decision-making and to improve the operation of the platforms themselves. In particular, as citizen reports become increasingly popular, there is a need to handle them more efficiently. In this paper, we present an analysis of ZüriWieNeu, a map-based website helping people in Zurich, Switzerland to report urban issues related to waste, broken streetlamps, or graffiti, among others. Our contributions are two-fold. First, we analyze what machine-extracted textual, visual, spatial and temporal features reveal about the dynamics of reporting and the content of each report category. This analysis provides a snapshot of the common patterns of urban issues in the Zurich area. Second, we perform classification to automatically infer the category of reports, achieving promising performance. Our work contributes towards developing machine learning-based systems to classify report categories, with the ultimate goal of supporting both users and platform operation.

CCS CONCEPTS

• **Mathematics of computing → Exploratory data analysis**; • **Computing methodologies → Supervised learning by classification**; *Natural language processing*; *Image segmentation*; • **Applied computing → E-government**.

KEYWORDS

civic computing; urban data; user-generated content; citizen reports; semantic segmentation; text embedding

ACM Reference Format:
Marcel Granero-Moya, Thanh-Trung Phan, and Daniel Gatica-Perez. 2021. Zurich Like New: Analyzing Open Urban Multimodal Data. In *Proceedings of the 1st International Workshop on Multimedia Computing for Urban Data (UrbanMM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3475721.3484310

UrbanMM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8669-2/21/10…$15.00
https://doi.org/10.1145/3475721.3484310

1 INTRODUCTION

Public administrations can improve the efficiency and accountability for damage in local infrastructures with citizen-driven technologies, as pointed out by Sturmer et al. [17]. Several online platforms have been created to efficiently collect citizen reports about problems in city's infrastructure. In 2007, FixMyStreet[1] was the first platform of this kind to be published. It allowed citizens in the UK to report problems such as a broken bench, confusing signalization, or offensive graffiti. The platform was built by mySociety, a non-profit social company, which made the source code available to the community. In 2013, the city of Zurich created its own FixMyStreet-based platform: "Züri wie Neu" (Zurich like new). Zurich's citizens can report issues via the phone application or the ZüriWieNeu's platform[2]. Users report problems with a title and a description, the location of the issue, and an optional picture. Each report belongs to a category that determines the type of issue (e.g., Graffiti, Signalization, etc.), which is then manually assigned to the administration in charge. So far, ZuriWieNeu has collected tens of thousands of user reports.

King et al. [9] and Abu-Tayeh et al. [2] surveyed citizens about FixMyStreet and ZuriWieNeu respectively. These open government platforms are well regarded by the citizens surveyed, as they feel empowered by bringing good to their local community. Citizens report increasing their quality of life and their engagement with the local community. Also, city governments are interested in exploiting these citizen-driven platforms to gain popularity and better handle damages to the city infrastructure. Hence, reporting systems such as ZuriWieNeu are likely to become more popular and be implemented more often in the future, and there will be the need to efficiently manage them.

In this work, we analyze ZüriWieNeu and share insights from this open urban dataset. In particular, we aim to provide ideas for data-driven systems for better handling of citizen reports. Reports are multimodal as they contain several information types: temporal (time and date), spatial (location of the issue), textual (title and description), and visual (photo). We use these feature modalities to train classifiers (with deep learning-based feature extractors), and analyze the classifiers to interpret the underlying nature of reports for each category. We address two research questions:

RQ1: What are the dynamics and content of citizen reporting in Zurich? How do reports vary over time and space? Which are the most common visual (i.e., semantic segmentation classes) and textual (i.e., words) features per category?

[1] FixMyStreet website: www.fixmystreet.com
[2] ZüriWieNeu's platform: https://www.zueriwieneu.ch/

RQ2: Can multimodal features be used to efficiently handle citizen reports by classifying them into categories? Which features are more relevant for this task?

The paper is structured as follows. In Section 2, we discuss related work. In Section 3, we describe the ZüriWieNeu dataset, and explain the data preprocessing steps. In Section 4, we address RQ1 by analyzing the dataset and describing each of its feature modalities. In Section 5, we address RQ2 by conducting a series of report classification experiments. In Section 6, we discuss how our work could benefit report handling, and also clarify potential concerns. Finally, we conclude the study in Section 7.

2 RELATED WORK

Citizen Reporting Systems. Urbanization will increase from 55% of the world's population today to 68% by 2050 [12]. As more people live in cities, authorities and citizens will face more issues related to public infrastructure [4, 6]. Local governments have natural bounds in terms of budget and maintenance staff to detect all types of issues occurring in cities. Fortunately, such problems can be reported by people directly to the proper city office [11]. Furthermore, local communities tend to know their urban environments well [15]. Hence, citizens can take part in collaborative systems to monitor their local infrastructure [10, 18]. A number of organizations and authorities have deployed mobile or web applications to encourage citizens to report local civic issues, using a variety of incentives [7].

Analysis of FixMyStreet-based Systems. Several studies have been conducted around FixMyStreet-based systems. Pak et al. [13] studied sociodemographic inequalities in Brussels. The analysis of FixMyStreet in this major European hub demonstrated how participation varies across the city's districts. Pak et al. discussed the need to improve this kind of reporting platform, with the specific goal of addressing the unintended marginalization of some communities, e.g., populations with specific cultural backgrounds. As another example in the European context, Parsons [14] leveraged data of FixMyStreet to analyze reports about dog fouling across the UK, studying the reasons why people express complaints, whether they attach media to reports, and the location of reports. Specifically for the city of Zurich, Abu-Tayeh et al. [1] identified reasons to use ZüriWieNeu, also built from FixMyStreet. This work found that people report either because they are self-concerned about a concrete issue or because they want to support other fellow citizens or the authorities. The authors proposed a way to motivate participants to contribute more reports, but alerted of a potential risk of leaving individuals with fewer digital skills behind, thus echoing the recommendations by Pak et al. [13] in the Brussels case. The above works used methods like surveys and quantitative data analysis to examine FixMyStreet or ZüriWieNeu.

Multimodal analysis of citizen reports. Citizen reports contain time, location, text descriptions, and photos. Some works have studied the application of machine learning methods on these multimodal data sources to characterize reported issues, ranging from heavy traffic [8] to so-called urban micro-events [19]. Santani et al. [8] studied Ma3Route, a mobile social media channel to crowdsource traffic reports in the city of Nairobi. Sukel et al. [19] used over 500K citizen reports from the city of Amsterdam and implemented a classification task of urban micro-events, e.g., loud boat, bicycle wreckage, etc. Sukel et al. proposed a real-time system, called Urban Object Detection Kit, to collect and analyze street-level images [20]. In this paper, we add to this literature by following a similar methodology and contribute a full multimodal analysis of ZüriWieNeu with data spanning over seven years, using both descriptive analyses and machine learning applied on contextual cues, text, and images.

Category	Reports	Media
Waste/CollectionPoint	7234	3739
Street/Sidewalk/Square	5441	2127
Signalization/LightSignal	4452	3385
GreenAreas/Playgrounds	2443	1825
Lighting/Clocks	2231	866
Graffiti	1330	1153
Other*	1511	946
Vbz/Ov** (public transport)	566	382
Wells/Hydrants**	384	89
General**	306	245
Pests**	255	230
Total	24642	14041

Table 1: Number of reports, and reports with media per category. Categories marked with ** are gathered into the meta-category Other*.

3 DATASET

3.1 ZurieWieNeu Data

"Züri wie neu", which stands for *Zurich like new*, is an online platform through which Zurich residents can report any imperfection in the urban infrastructure. The city of Zurich shared an open dataset[3] with reports dating from May 2013 to October 2020. Overall there are 24642 reports, from which 14041 contain media. This dataset is interesting because of its multimodal nature. Each report contains the time when it was posted, the location of the issue, a title and a description, an optional URL pointing to the uploaded image, and a category which represents the responsible department for that issue (see Table 1 for the distribution of categories.)

3.2 Data preprocessing

Data was preprocessed to be understandable and prepared for classification. Firstly, categories containing the lowest number of reports were gathered into a new meta-category, *Other*, to make the dataset more balanced. Table 1 shows the total number of reported issues and the number of reports containing media per category. Secondly, location data was standardized by mapping it from the Swiss reference system CH1903+_LV95 to the global standard WGS84. We used a transformation script[4] provided by the Swiss Federal Office of Topography. Thirdly, we downloaded all media using the dataset URLs and obtaining color JPEG images of around 188×250 pixels. Due to the small size, all media could be stored in 195.5 MB.

Finally, textual information was provided in Swiss German or High German. Given the few resources for these languages, such

[3]The dataset is available in openswiss.data website.
[4]github.com/ValentinMinder/Swisstopo-WGS84-LV03/wgs84_ch1903.py

as standardized text corpora and pre-trained embeddings, we translated titles and descriptions into English to use state-of-the-art NLP models that work for English. Text was translated with the Neural Machine Translation (NMT) model provided by the Google Translation API. Note that Swiss German is considered a German dialect as it lacks a standard written form, which may complicate the translation and degrade text quality. Apart from translation, category keywords were removed from descriptions. An example of the text preprocessing steps appears in Table 2.

Example of Reported Issue	
Time & date	09:40:23 09/04/2013
Location	(latitude, longitude)
Source text	Schaukel Spielplatz Stolzewiese: Schaukel fehlt auf dem Spielplatz Stolzewiese
Translated text	Swing at the Stolzewiese playground: There is no swing at the Stolzewiese playground
Translated text w/o keywords	Swing at the Stolzewiese: There is no swing at the Stolzewiese
Image (optional)	
Category	GreenAreas/Playgrounds
Platform URL	https://www.zueriwieneu.ch/report/41

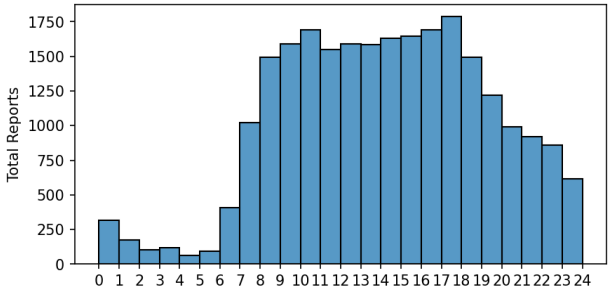
Table 2: Example of report containing the temporal, spatial, textual and visual information, as well as the category and a link to the platform. The textual information shows the normalization steps for translation and keyword removal.

4 RQ1: ANALYSIS OF FEATURES BY MODALITY

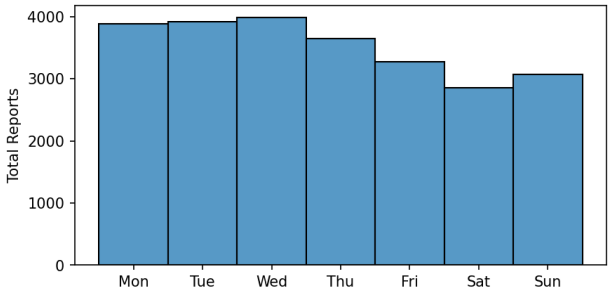
In this section, we analyze the dataset, describing feature by feature and illustrating with examples. First, contextual features (i.e., time and location) are analyzed. Second, textual and visual information are transformed into denser, more meaningful features. These features are extracted from text and images through open pre-trained machine learning models: text is embedded into sentence vectors, and semantic segmentation is applied to images. These features are analyzed and used to describe report categories.

4.1 Temporal patterns

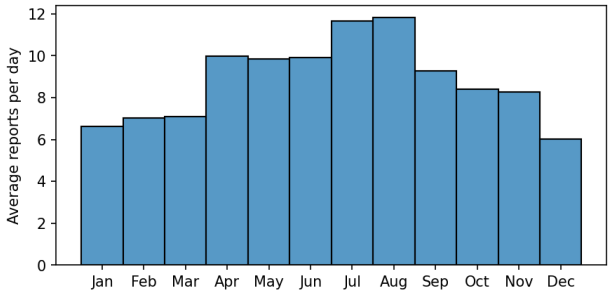
In our ZuriWieNeu dataset, reports are dated from 2013-03-14 to 2020-10-03. By examining the yearly amount of reports, we notice a yearly increase of reports starting from 2015 until 2019. In contrast, 2020 does not follow this increasing trend: it has a decrease of reports because only 9 months were recorded and also due to the Covid-19 pandemic. Figure 1a depicts the hourly distribution when reports were posted. Reports are equally recorded during working hours (8am-6pm), and then the number of reports decreases. Sleeping hours register low activity. Regarding weekday distribution in Figure 1b, issues are equally reported throughout the week with a decrease on Friday, Saturday, and Sunday. Finally, Figure 1c illustrates the monthly distribution of reports. Issues are reported in the platform during the whole year, but from April to September more activity is recorded. We hypothesize that the increase in reports is caused by an increase of outdoors activity for Zurich citizens during these months, with more daylight hours and better weather.



(a) Total reports per hour



(b) Total reports per weekday



(c) Average of daily reports per month

Figure 1: Histograms for temporal features

4.2 Location features

The City of Zürich is divided into 12 districts and 34 quarters. We used the quarters division[5] to analyze the location of reports. Figure

[5]Geojson downloaded from github.com/blackmad/neighborhoods

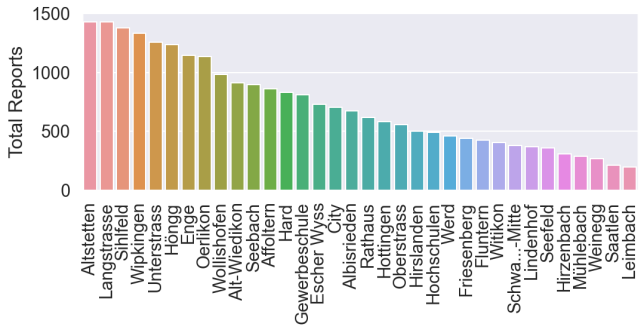


Figure 2: Histogram of reports per quarter.

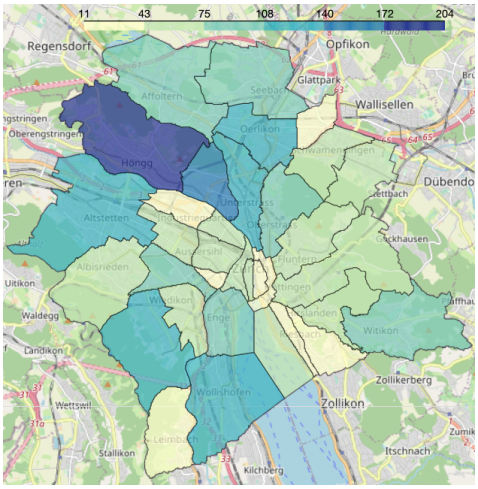


Figure 3: Heatmap of reports for *GreenAreas/Playgrounds*. Yellow indicates a lower number of reports, Green indicates a higher number, and Blue the highest number.

2 shows a ranked histogram of reports per quarter. The 6 quarters with the greatest amount of reports is located in western and central areas of Zurich.

An example of a specific report category, shown in Figure 3, indicates that issues concerning green areas and playgrounds are reported in less central areas. As a matter of fact, the quarter of Höngg, located at the western border of Zurich, is the area with the highest number of reports as two big forests are located in it.

4.3 Textual descriptions

We embedded the translated descriptions without category keywords into vectors for later classification. We used InferSent[6] [5], an English sentence encoder which provides semantic embeddings in 4096-dimension vectors. Authors claim that it generalizes well to many natural language processing tasks. Before applying InferSent to the descriptions, the names of report classes were removed from descriptions to make a fairer classification between classes. After using the sentence encoder, the result is a vector of 4096 embedded dimensions for each report.

We analyze the most frequent words to describe the most common issues per category. Note that, as explained in Section 3, category keywords were removed. Figure 4 shows the 10 most common words per category, and in the last plot, it shows the most frequent words for the whole dataset.

The most frequent words give insights on the issues reported in each category. For instance, in the Graffiti category the most common words are *wall* and *sprayed*, which actually define what graffiti is. The list also contains *smear* and *racist*, which reveals the reason users considered them an issue worth reporting. In Lightning/Clocks, *lamp* appears 8% of the times, and as users report issues, words like *defective* and *dark* show that the lighting was not working, and it had to be fixed. For GreenAreas/Playgrounds, the most frequent words are related to *trees* and *broken* paths or benches. Signalization/LightSignal category reports issues related to traffic, pedestrian zones and signs. Overall, all categories contain coherent words related to the category, and words reporting why it was an issue such as *illegal*, *broken*, or *defective*. The meta-category Other contains *mosquito* and *tiger* because one of its original categories was Pests, and many reports were about the Asian tiger mosquito.

4.4 Image content

We performed Semantic Segmentation of images with a Resnet50 pre-trained[7] on the ADE20K dataset [21, 22]. The result of the semantic segmentation was a mask mapping the predicted semantic class for each pixel in the image, as shown in Table 3. To reduce the dimensionality of this feature, the mask was flattened into a 150-dimension vector. This feature vector represents the percentage of pixels classified into each of the 150 semantic visual features defined by the ADE20K dataset. In the example of Table 3, the original image (left) is segmented semantically into a mask (right). The mask is used to compute the percentage of pixels per semantic class. As there are 150 semantic classes, the resulting vector will contain zeros in all dimensions, but the 5 classes detected, i.e., tree, signboard, sky, building, and plant.

Figure 5 depicts the most common visual features per report class. In the x-axis, the 11 most common visual features are shown for each of the report classes. These 11 most common features are chosen by aggregating the top-8 visual features per report class. Therefore, images from different report classes also share many semantic visual features. This lack of diversity can be explained by the context of the dataset. The data consists of urban user-generated images that try to report an issue in the city; thus, the most common features are strongly linked to an outdoor urban landscape such as roads, buildings, sidewalks, or signboards.

The category for Green Areas/Playgrounds has the highest scores for the features related to nature, i.e., tree, earth;ground, plant;flora, and grass. Similarly, Graffiti category photos get the highest scores for building;edifice and wall, as it is where graffiti are painted. Regarding Lightning/Clocks and Signalization/LightSignal, both get the highest scores in sky, these can be interpreted by the fact that lamps, clocks, and signals are in more elevated positions with

[6]InferSent GitHub page: github.com/facebookresearch/InferSent

[7]Instructions for the ResNet50 pre-trained on ADE20K: cv.gluon.ai/build/ examples_segmentation/train_psp

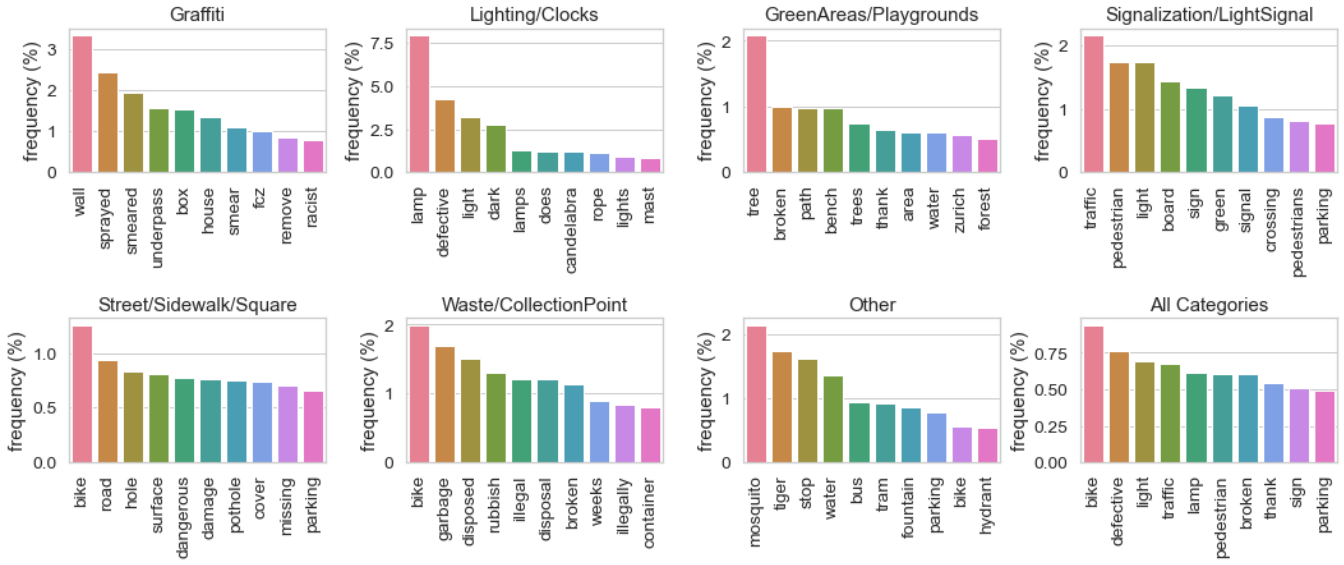
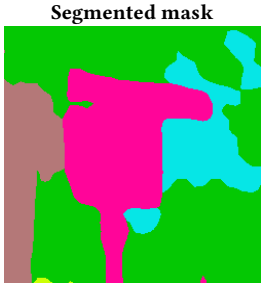


Figure 4: Most frequent words per report category and for all categories.

Original image

Segmented mask



Semantic Class (ADE20K)	Color	Percentage
tree	green	27.079
signboard;sign	pink	22.587
sky	sky blue	21.909
building;edifice	brown	19.262
plant;flora;plant;life	yellow	6.345

Table 3: Example of semantic segmentation. Original image at top left, semantic segmentation mask at top right, and percentage of pixels per semantic class at bottom. Source: image 20937.

respect to the ground and users have to take pictures pointing upwards. Both categories obtain similar values for all features but road;route and signboard; it is coherent because traffic lights and signals are built in roads. Finally, Street/Sidewalk/Squares category refers to issues like potholes or missing paving stones, therefore it is reasonable for this category to have the highest values for road;route and sidewalk;pavement features.

5 RQ2: CLASSIFICATION OF REPORTS

In this section, we leverage the multimodal features described in Section 4 to classify reports in categories. As explained in Section 3, there are 7 report categories, including *Other*. We perform two types of classification: (1) category classification, identifying to which of the 7 possible categories a report belongs to; and (2) binary classification, detecting whether a report belongs to a given category or not. Random forest classifiers [3] with 500 trees are used in both classification types. The metric used to evaluate the classifiers performance will be the mean class accuracy, i.e., the average of classes accuracy. After each experiment, we show feature importance depending on Gini importance (or mean decrease impurity) to highlight the most relevant features and their typology. Table 4 indicates the dimensions for each feature type.

Time (5)	hour (1), day (1), month (1), weekday (1), year (1)
Location (36)	latitude (1), longitude (1) and quarter (34)
Image (150)	percentage of pixels per semantic class (150)
Text (4096)	sentence embedding (4096)

Table 4: Multi-modal features used for classification. The dimension of each feature is shown between parentheses.

5.1 Category classification based on features type

The goal of this experiment is two-fold, namely to obtain the best category classifier for the 7 categories, and to perform an ablation study about the best feature types for category classification. As shown in Table 5, 15 random forest classifiers are fit with all the possible combinations of feature types.

The accuracy scores for the experiment are shown in Table 5. We use the majority class (*Waste/CollectionPoint*) as the accuracy baseline (27.8%). Then, classifiers are sorted per accuracy. All feature types alone outperform the baseline, which means that they provide information by themselves. When comparing them, the classifier

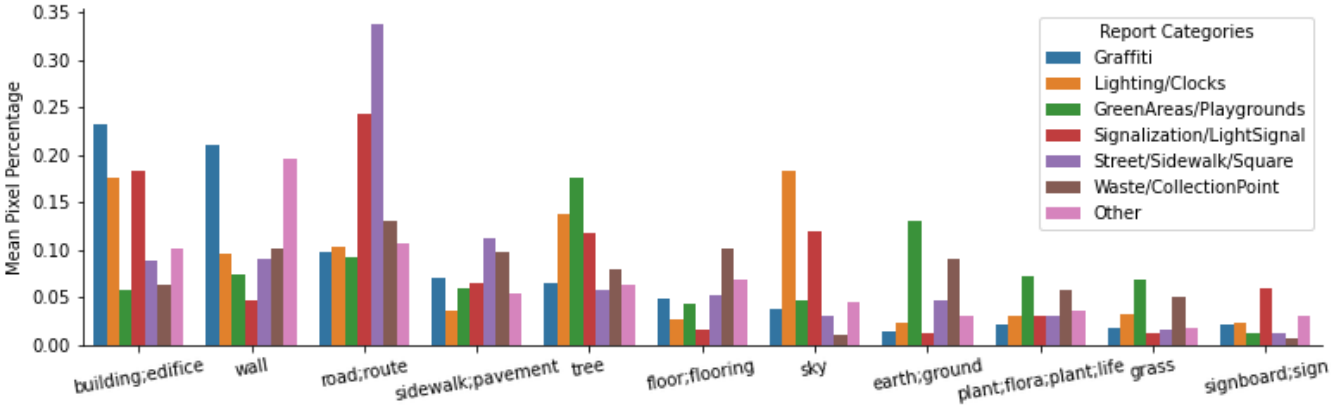


Figure 5: Most frequent visual features. Mean pixel percentage of the most common semantic features shown per report class

Text	Image	Location	Time	Accuracy
Baseline (Majority class)				27.8%
			X	29.8%
		X		37.8%
	X			54.3%
X				**66.5**%
		X	X	40.7%
	X		X	56.2%
	X	X		56.3%
X		X		66.5%
X			X	66.8%
X	X			**67.9**%
	X	X	X	58.3%
X		X	X	67.1%
X	X	X		68.4%
X	X		X	**68.5%**
X	X	X	X	**67.8**%

Table 5: Results for category classification depending on the type of features. The top row is for the majority class baseline. The rest of rows show the accuracy for the random forest when using the features marked with an *X*.

trained with time features barely outperforms the baseline (29.8%), it improves to 37.8% when using location features, image features reach a 54.3% accuracy, and text outperforms the other feature types (66.5%) and almost achieving the maximum accuracy. When combining feature types, classifiers using text features outperform classifiers without text features. By splitting scenarios with and without text, models trained with image features improve over models without them. Ultimately, we can state that text features are the most important ones followed by image features; and location and time features make a small contribution. The classifier reaches its top result with 68.5% of mean class accuracy.

For further interpretability, we show the most important features per classifier in Figure 6. The classifier trained with time features alone shows that *hour* is the most important time feature followed by *day*, *month*, *weekday* and *year* respectively. For location features, *longitude* and *location* are equally significant, whereas the one-hot-encoded quarter features have almost no importance. Regarding the random forest trained on visual features, the most common visual features, as in Figure 5, are the most important features for the classifier. In contrast, textual features are difficult to interpret because they are named with the number of the embedding dimension extracted by InferSent. Nevertheless, textual features can be interpreted by observing the report descriptions with the highest and lowest values. For instance, text feature *1096* obtains high values for descriptions referring to candelabra, a type of streetlamp.And another feature, *46*, is triggered for descriptions mentioning traffic or pedestrian lights.

When combining both textual and visual features, the classifier relies on both types of features, which suggests that image and text complement each other. The classifier trained with all feature modalities mainly relies on visual and textual features, except for *year* which is the 6th most relevant feature, and *longitude* and *latitude* which are also among the top-15 features.

5.2 Binary Classification of reports

The second experiment trains the random forest classifiers for a binary task with all feature types combined. The classifier has to detect whether reports belong to a given category or the rest of the categories. Each random forest is trained with a subset of the dataset containing the set of reports (with media) for the given category, and an equally-sized set with reports randomly selected from the rest of categories. The goal of this task is to identify categories that are more challenging than others. The resulting accuracies are depicted in Table 6. The lowest performance is obtained for the classifier for *Other* meta-category, and the best result is obtained for the classifier for *Signalization/LightSignal*.

Figure 7 shows the feature importance for the binary classification of each category. These plots show which features are more relevant to determine the classification result. *Graffiti* and *Lighting/Clocks* classifiers mainly base their decision on text features. This fact matches with the corresponding confusion matrices (not

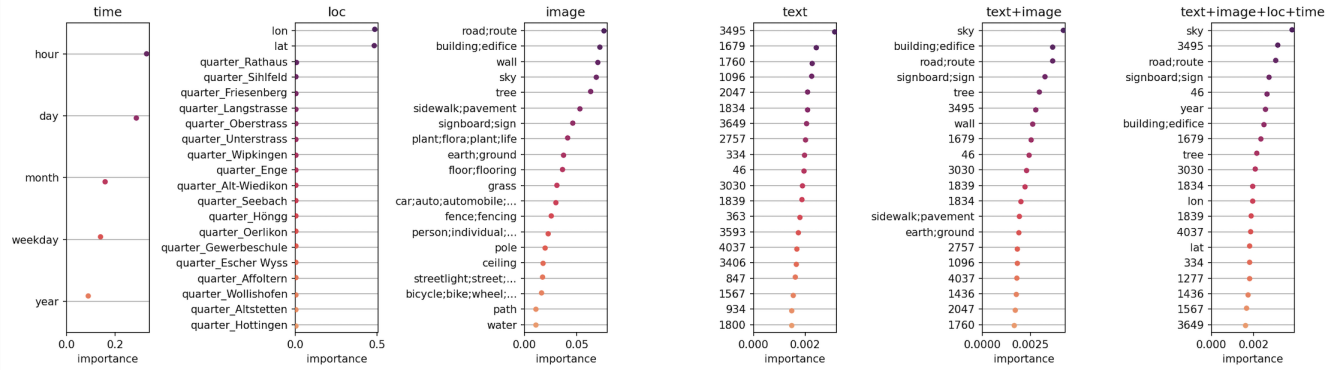


Figure 6: The 20 most important features of category classification for main combinations of feature types.

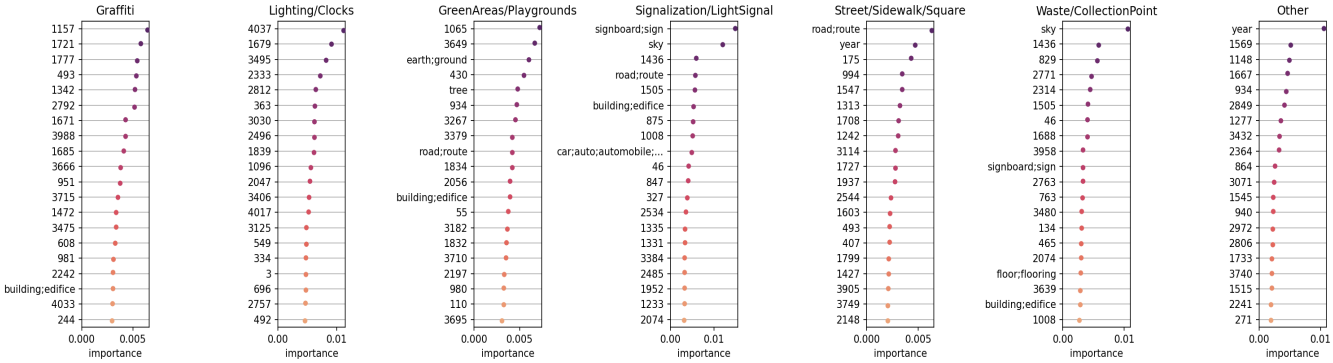


Figure 7: The 20 most important features for binary classification for each category.

Category	Samples	Accuracy
Baseline (Majority class)		50.0%
Other	1892	75.2%
Street/Sidewalk/Square	4254	78.5%
GreenAreas/Playgrounds	3650	80.4%
Graffiti	2306	85.9%
Waste/CollectionPoint	7478	86.8%
Lighting/Clocks	1732	87.3%
Signalization/LightSignal	6770	**89.0%**

Table 6: Binary classification of report categories using all type of features (time, location, image, and text). The number of training samples are specified in the second column.

shown for space reasons.) The accuracies of both categories increase in the model based on text compared to the model based on visual features. Therefore, a binary classifier trained for these categories mainly relies on text features. For the other four categories in the figure, the most important visual features usually are the most common features of the target category. Time and location features are not relevant except for the year feature in categories *Street/Sidewalk/Square* and *Other* due to their different yearly distribution.

6 DISCUSSION

We now discuss the implications of the results of our study.

Potential utility of learned models. While the performance of the classifiers can certainly be improved, we see potential value for this approach. The trained models could improve the Zuri-WieNeu online platform in two different ways. First, the multi-class classifier could be used to recommend a category to users once they have written a report, and thus facilitate this part of the report generation. On the other hand, the binary classification, which reaches higher accuracy, could be applied to show a warning if the category selected by the user does not match the binary classifier prediction for that category. These methods have the potential to increase efficiency when managing citizen reports.

Privacy issues. During the study, we were concerned about privacy, as potentially recognizable faces may accidentally appear in the images as citizens take photos. In practice, we detected a number of pictures that contained a semantic visual feature related to people, more specifically 'person;individual;someone; somebody;mortal;soul'. To assess the privacy risks in the dataset, we manually evaluated a set of 100 random images out of the 1838 images containing that visual label. We found that 21 images had people on them, and that 2 photos showed potentially recognizable faces. On one hand, this result appears logical given that the objective of citizens is to photograph places and infrastructure, some

of which are above the ground (e.g., traffics signs) and thus not likely to capture people. On the other hand, this might also suggest that citizens are being mindful of not including passersby in their reports. Further qualitative research based on interviews with contributors to the platform could shed light on these hypotheses.

Applicability to other cities. These methods were used for citizen reports in Zurich, but we believe that they are scalable to other cities, as also shown by recent work [19]. We argue that location and time can be collected similarly, and text can be translated to English, as in this work. On the other hand, the ResNet50 that extracts the semantic visual features is likely to contain bias towards western countries, as is the case with several the popular image databases used for deep learning [16]. This is a problem recently illustrated in [8] in the context of African cities. This situation calls for careful thinking about what parts of the methodology would have to be adapted or retrained to capture the large diversity in cities, specifically in the Global South.

7 CONCLUSION

This work presented an analysis of ZüriWieNeu, an open urban dataset about user reports of damage in the city of Zurich; and showed how these reports can be automatically classified into categories representing the type of issue. Each report contains multimodal information (time, location, text description, and an optional picture), and is tagged with a category from a rich number of possible issues.

Regarding our RQ1, we analyzed each of the multi-modal features to understand the nature of the dataset. We concluded that temporal patterns follow expected daily rhythms, and that more issues are reported during the spring and summer seasons, likely due to the increase in outdoor activity. Location features show that the majority of issues are reported in central and western quarters of the city, and that the spatial distribution changes depending on the report category. Regarding textual descriptions, the most frequent words vary strongly depending on the category and describe the typology of issues. Finally, the most frequent visual semantic classes showcase an outdoor urban landscape and coherently change among categories.

Regarding our RQ2, two different experiments with random forest classifiers leveraged the features to classify reports with promising performance. The best classifier of the first experiment reached 68.5% of accuracy; it inferred the report's category out of seven possible classes. In the second experiment, classifiers inferred whether a report belongs to a category or not. Half of the categories achieved a test accuracy of over 85%. We conclude that multimodal features can be used to handle citizen reports by classifying them into categories, while clearly there is scope for performance improvement. Finally, feature importance was analyzed in both experiments to provide some level of explainability, and this analysis concluded that the most important features are textual and visual, in decreasing order. This confirms that textual description and photos in citizen reporting platforms are both important.

ACKNOWLEDGMENTS

We thank the city of Zurich for the open Züri wie neu dataset. D. Gatica-Perez was partly supported by the IcARUS project, funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 882749.

REFERENCES

[1] Gabriel Abu-Tayeh, Oliver Neumann, and Matthias Stuermer. 2018. Exploring the motives of citizen reporting engagement: Self-concern and other-orientation. *Business & information systems engineering* 60, 3 (2018), 215–226.
[2] Gabriel Abu-Tayeh, Edy Portmann, and Matthias Stürmer. 2017. „Züri wie neu ": Public Value von Online-Partizipation. *HMD Praxis der Wirtschaftsinformatik* 54, 4 (2017), 530–543.
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Ellen M Brennan and Harry W Richardson. 1989. Asian megacity characteristics, problems, and policies. *International Regional Science Review* 12, 2 (1989), 117–129.
[5] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).
[6] Diogo Costa, Paolo Burlando, and Cindy Priadi. 2016. The importance of integrated solutions to flooding and water quality problems in the tropical megacity of Jakarta. *Sustainable Cities and Society* 20 (2016), 199–209.
[7] David N Crowley, John G Breslin, Peter Corcoran, and Karen Young. 2012. Gamification of citizen sensing through mobile social reporting. In *2012 IEEE International Games Innovation Conference*. IEEE, 1–5.
[8] Daniel Gatica-Perez, Darshan Santani, Joan Isaac-Biel, and Thanh-Trung Phan. 2019. Social Multimedia, Diversity, and Global South Cities: A Double Blind Side. In *Proceedings of the 1st International Workshop on Fairness, Accountability, and Transparency in MultiMedia*. 4–10.
[9] Stephen F King and Paul Brown. 2007. Fix my street or else: using the internet to voice local public service concerns. In *Proceedings of the 1st international conference on Theory and practice of electronic governance*. 72–80.
[10] Fernando Koch, Carlos Cardonha, Jan Marcel Gentil, and Sergio Borger. 2012. A platform for citizen sensing in sentient cities. In *International Workshop on Citizen in Sensor Networks*. Springer, 57–66.
[11] Hana Kopackova and Petra Libalova. 2019. Quality of citizen reporting tools at municipal level. *Journal of Information Systems Engineering and Management* 4, 3 (2019), em0092.
[12] United Nations. 2018. *68% of the world population projected to live in urban areas by 2050, says UN*. Retrieved July 23, 2021 from https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html
[13] Burak Pak, Alvin Chua, and Andrew Vande Moere. 2017. FixMyStreet Brussels: socio-demographic inequality in crowdsourced civic participation. *Journal of Urban Technology* 24, 2 (2017), 65–87.
[14] Alex Parsons. 2017. Dog Fouling and FixMyStreet. *mySociety* (2017). https://research.mysociety.org/media/outputs/fms_dog_fouling.pdf
[15] Darshan Santani, Salvador Ruiz-Correa, and Daniel Gatica-Perez. 2018. Looking south: Learning urban perception in developing cities. *ACM Transactions on Social Computing* 1, 3 (2018), 1–23.
[16] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. 2017. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. In *NIPS 2017 Workshop on Machine Learning for the Developing World*.
[17] Matthias Stürmer, Oliver Neumann, and Tim Loosli. 2017. Heaven for grumblers or the road to better public services? A study of critical reports from a Swiss citizen co-production system. In *Proceedings 21st Annual International Research Society for Public Management conference (IRSPM)*.
[18] George Suciu, Lucian-Alexandru Necula, Vladimir Jelea, Dragos-Sebastian Cristea, Carmen-Catalina Rusu, Luigi-Renato Mistodie, and Marius-Petrut Ivanov. 2021. Smart City Platform Based on Citizen Reporting Services. In *Advances in Industrial Internet of Things, Engineering and Management*. Springer, 87–100.
[19] Maarten Sukel, Stevan Rudinac, and Marcel Worring. 2019. Multimodal classification of urban micro-events. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1455–1463.
[20] Maarten Sukel, Stevan Rudinac, and Marcel Worring. 2020. Urban Object Detection Kit: A System for Collection and Analysis of Street-Level Imagery. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 509–516.
[21] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.
[22] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* 127, 3 (2019), 302–321.