Multilingual Sentiment and Topic Analytics for FixMyStreet Brussels: Spatio-Temporal Hotspot Detection and Decision Support from Citizen Reports

Marian Pompiliu Cristescu *

Posted Date: 15 April 2026

doi: 10.20944/preprints202604.1116.v1

Keywords: citizen sensing; FixMyStreet; multilingual NLP; urgency classification; topic modeling; hotspot detection; calibration; explainable AI

Article

Multilingual Sentiment and Topic Analytics for FixMyStreet Brussels: Spatio-Temporal Hotspot Detection and Decision Support from Citizen Reports

Marian Pompiliu Cristescu

Lucian Blaga University of Sibiu, Calea Dumbrăvii, no. 17, 550324, Sibiu, Romania;
marian.cristescu@ulbsibiu.ro

Abstract

Citizen-reporting platforms generate high-volume, multilingual streams of service requests, yet operational triage often relies on coarse category labels and manual inspection. This study develops an explainable, calibration-aware analytics pipeline for FixMyStreet Brussels reports, combining text-based urgency modeling, topic discovery, and spatio-temporal hotspot scoring to support municipal decision-making. From 522,132 raw reports, we build an English-normalized text field for modeling, derive resolution-time outcomes from closed cases, and curate a 1,000-item gold standard with an explicit high-urgency class. A TF–IDF logistic regression baseline achieves strong classification performance and, after probability calibration, yields well-behaved confidence estimates suitable for risk-aware prioritization. Topic-level analyses reveal dominant themes related to sidewalks, road damage, and bulky waste, and hotspot scores highlight persistent, high-impact issue clusters. Event detection on aggregated signals did not identify statistically significant shocks during the analysis window, suggesting that the observed dynamics are driven by chronic, recurring problems rather than abrupt anomalies. Explainability audits via SHAP expose linguistically intuitive drivers for urgent cases (e.g., dangerous, risk, accident) and complaint-oriented terms (e.g., abandoned, illegal, dirty), providing transparent hooks for governance review.

Keywords: citizen sensing; FixMyStreet; multilingual NLP; urgency classification; topic modeling; hotspot detection; calibration; explainable AI

1. Introduction

Digital civic-participation tools have expanded the role of citizens in urban service monitoring, enabling residents to report local problems such as illegal dumping, damaged sidewalks, or lighting failures at fine spatial and temporal granularity. These reports constitute a form of volunteered geographic information and can complement official sensing and inspection regimes by increasing coverage and responsiveness [1] (pp. 211–221), [2] (pp. 682–703). At the same time, high submission volumes and heterogeneous free text create operational bottlenecks: municipal teams must triage competing requests, manage backlogs, and identify emerging trouble spots while maintaining accountability. Decision-support methods that integrate language signals with spatio-temporal patterns are therefore increasingly relevant for public administrations seeking data-driven service allocation.

FixMyStreet platforms provide a useful testbed for such methods because they mix structured fields (category, status, timestamps, location) with citizen narratives that often encode urgency cues and situational context. However, text inputs are multilingual, noisy, and frequently short; labels may reflect administrative routing rather than risk; and predicted "priority" scores must be interpretable and properly calibrated to be actionable. The goal of this work is to deliver a transparent analytics pipeline that (i) models report urgency from text in a way that supports risk-aware ranking,

 doi:10.20944/preprints202604.1116.v1

(ii) summarizes recurrent problem themes via topic models, and (iii) quantifies spatio-temporal hotspots and potential event-like deviations using simple, auditable statistics.

2. Literature Review

Research on citizen reporting and urban informatics has emphasized the value of crowdsourced observations for municipal services, planning, and crisis response [1] (pp. 211–221). Empirical work on volunteered geographic information has also highlighted systematic quality challenges—uneven participation, positional uncertainty, and reporting bias—which motivate methods that combine human oversight with automated screening [2] (pp. 682–703).

Natural language processing has been widely used to extract sentiment, complaints, and urgency from short texts, including social media and customer-feedback streams. In public-service contexts, urgency signals are often expressed through hazard and safety language (e.g., danger, accident) and through negative affect related to persistence or recurrence (e.g., again, always). Transformer architectures such as BERT have improved semantic representations for downstream classification tasks [3], and multilingual pretraining (e.g., XLM-R) supports cross-lingual transfer where labeled data are sparse [4]. Nevertheless, linear baselines remain attractive in governance settings due to their transparency and ease of audit, especially when paired with post-hoc explanations.

Topic modeling is commonly used to organize large text corpora into interpretable themes. Classical probabilistic approaches such as latent Dirichlet allocation (LDA) provide a principled generative framework [5] (pp. 993–1022), while newer neural topic models combine contextual embeddings with class-based term weighting to improve coherence on short texts [6]. For municipal operations, topic models can surface recurring issue types, connect them to administrative categories, and facilitate targeted interventions.

Spatial and spatio-temporal hotspot analysis has a long history in geography and epidemiology. Local indicators of spatial association and scan statistics are frequently used to detect clusters and prioritize field action [7] (pp. 93–115), [8] (pp. 1481–1496). When operational decision-making requires transparent heuristics, z-score based anomaly indices and simple aggregation rules can offer useful, interpretable signals—particularly when combined with language-derived severity probabilities. Event detection in time series has also been extensively studied, including change-point methods that identify distributional shifts or mean changes [9].

Finally, deploying predictive models in public administration requires attention to reliability and accountability. Probability calibration is essential when model scores are used as risk estimates rather than as mere rankings. Miscalibrated models can mislead triage, either by overstating certainty or by underestimating rare but critical risks. Post-hoc calibration methods and reliability diagrams are standard tools for evaluating probabilistic correctness [10,11]. Explainability methods such as SHAP provide feature-level attributions compatible with linear models and can support governance audits [11].

2. Materials and Methods

The analysis uses an export of FixMyStreet Brussels reports as seen in Table 1, containing identifiers, timestamps, hierarchical categories, responsible organizational units, status, address descriptors, free-text narratives, and point geometries. Resolution time is computed from created and closed timestamps for closed cases.

Table 1. Data schema overview for FixMyStreet Brussels reports.

Field(s)	Type	Description
FID, gid, fims_id	Identifier	Unique record identifiers for traceability.

 doi:10.20944/preprints202604.1116.v1

createddate, updateddate, closeddate	Temporal	Lifecycle timestamps for calculating resolution time and age.
category, head_category	Taxonomy	Hierarchical issue classification (e.g., Public Cleanliness).
responsible_org, responsible_dep	Administrative	Department responsible for intervention.
status	Operational	Current state (e.g., Open, Closed, Transferred).
road_fr, road_nl, pccp	Location	Address descriptors and postal codes.
comment, comment_reporter	Narrative	Free-text citizen report and source indicator.
comment_translated	Derived text	English-normalized text used for modeling.
geom	Spatial	Point geometry (EPSG:4326/31370).

Free-text comments are cleaned by removing obvious system templates and non-informative tokens, standardizing whitespace and line breaks, and filtering empty entries. A language-normalization step produces an English field (comment_translated) used for downstream modeling. Short texts (<10 characters) are retained but flagged for robustness checks, since they provide limited semantic evidence. Gold-standard annotation and label scheme. A stratified sample of 1,000 reports is manually labeled into three classes representing non-urgent/neutral (class 0), routine complaint/service request (class 1), and high urgency/safety risk (class 2). The high-urgency class is intentionally rare to reflect operational reality and to stress-test recall and calibration for critical cases. (see Table 2)

Table 2. Preprocessing and sampling statistics.

Metric	Value	Notes
Total raw reports	522,132	Initial ingestion
Exact duplicates	760	Collapsed based on timestamp + geometry + text
Closed reports	426,490	Used for resolution-time calculation
Empty/null comments	14.25%	Excluded from text modeling
Short text (<10 chars)	15.57%	Flagged for robustness checks
Gold standard size	1,000	Manually annotated subset
Gold: high urgency (class 2)	50	Safety risks / urgent items

A bag-of-words baseline is trained using TF–IDF features and multinomial logistic regression. Model quality is assessed with accuracy and macro-averaged F1. Because downstream decisions require interpretable probabilities, model confidence is evaluated via reliability diagrams and expected calibration error (ECE). A calibrated variant is obtained by fitting a post-hoc calibration model on held-out predictions. Confusion matrices support error analysis across classes.

To summarize recurrent problem types, reports are clustered into topics using a term-based topic representation that yields top words per topic. For each topic, the mean predicted urgency probability and an aggregate hotspot score are computed, enabling comparisons between high-volume but low-risk themes and lower-volume themes that concentrate risk.

Reports are aggregated over spatial units and weekly time bins. Hotspot scores combine normalized volume and urgency signals to highlight areas with persistent, high-impact demand. A z-score based event detector is applied to aggregated time series; events are flagged only when deviations exceed a predefined threshold, reducing sensitivity to routine seasonality and noise.

 doi:10.20944/preprints202604.1116.v1

For the TF–IDF logistic regression model, SHAP values are computed to identify terms that increase or decrease the predicted probability of high urgency and complaint-oriented classes. Term rankings are inspected for face validity and for potential governance concerns (e.g., artifacts, overly generic tokens).

3. Results

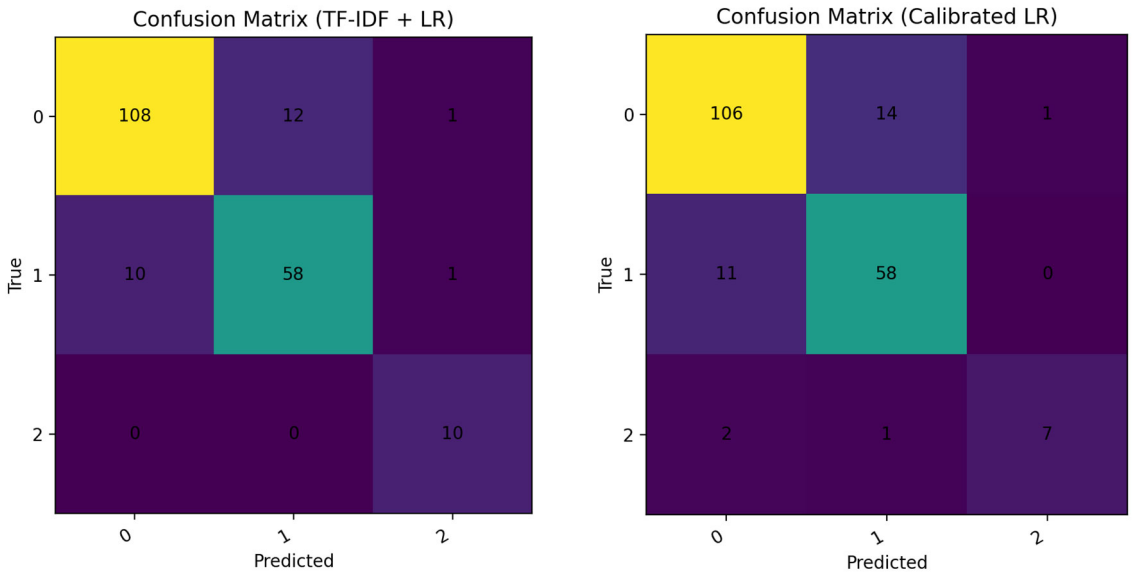
As seen in Table 2, the export contained 522,132 raw reports. Exact duplicates were rare (760 records) and were collapsed using timestamp, geometry, and text keys. Closed reports represented the majority of cases (426,490) and were used to compute resolution-time outcomes. Text fields were frequently sparse: 14.25% of reports had empty or null comments and were excluded from text modeling; 15.57% contained fewer than 10 characters and were retained as low-information cases.

Analyzing the results in Table 3 on the gold-standard test set we can identify that the calibrated TF–IDF logistic regression model achieved accuracy of 0.855 and macro F1 of 0.826, with a high-urgency (class 2) recall of 0.700. The uncalibrated model variant produced higher overall scores on this split (accuracy 0.880; macro F1 0.882) but exhibited poorer probability behavior in reliability analysis.

Table 3. Urgency classification performance on the gold-standard split. ECE for the calibrated model is 0.069.

Model	Accuracy	Macro F1	Urgency (class 2) recall
TF–IDF + LR (raw)	0.880	0.882	1.000
TF–IDF + LR (calibrated)	0.885	0.826	0.700

Confusion matrices (Figure 1) indicate that most errors occur between classes 0 and 1, consistent with overlap between neutral reporting and routine complaints. High-urgency cases are typically recognized, but calibration shifts probability mass toward more conservative predictions, increasing some misses of class 2 in exchange for improved probability trustworthiness.

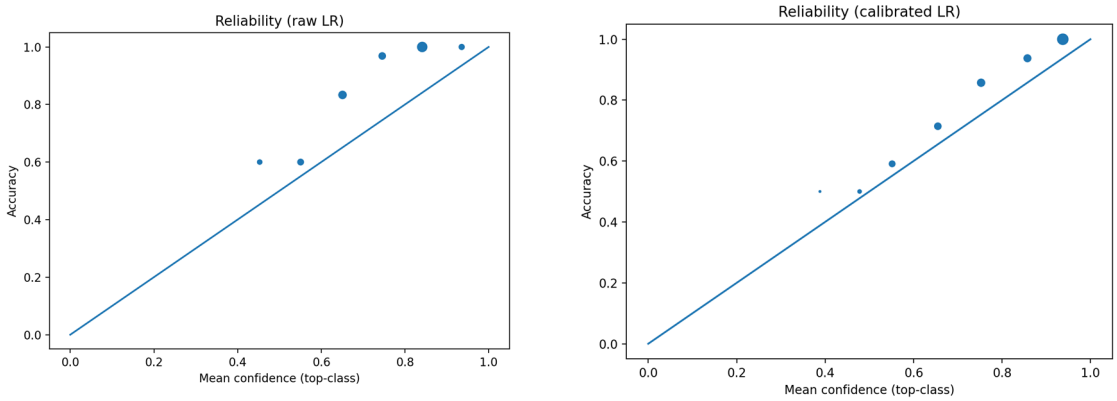


(**a**) (**b**)

Figure 1. Confusion matrix: (**a**) Raw TF–IDF logistic regression urgency classifier (3 classes). Values are counts on the gold-standard test split.; (**b**) Calibrated TF–IDF logistic regression urgency classifier. Calibration yields more conservative class-2 predictions on the same evaluation split.

 doi:10.20944/preprints202604.1116.v1

Reliability diagrams show that the calibrated model's mean confidence aligns more closely with empirical accuracy across bins, while the raw model tends to be overconfident at higher predicted probabilities, as seen in Figure 2. The calibrated model achieved ECE of 0.069, supporting the use of its probabilities as decision-support scores rather than only as rankings.



(**a**) (**b**)

Figure 2. Reliability diagram: (**a**) Raw TF–IDF logistic regression model. Points above the diagonal indicate overconfident probabilities in the corresponding bins; (**b**) Calibrated TF–IDF logistic regression model. Calibration improves alignment between mean confidence and empirical accuracy.

As seen in Table 4, topic modeling surfaced a small number of dominant operational themes. The largest topic reflected sidewalk and road-surface hazards (topic 14; volume 244,568), followed by templated acknowledgment or routing messages (topic 11; volume 144,160). Across topics, mean urgency probabilities were relatively low (approximately 0.02–0.09), consistent with the rarity of high-urgency labels. Hotspot scores differentiated themes that are both frequent and spatially concentrated. Topics linked to physical hazards and obstructions (e.g., sidewalk damage, bulky waste, fallen trees) tended to show elevated hotspot scores relative to purely administrative or templated text. Some high hotspot values were associated with noisy tokens (e.g., line-break artifacts), highlighting the need for continued text hygiene in production deployments.

Table 4. Topic–urgency–hotspot interaction (top topics by volume; top words abridged).

Topic	Top words (abridged)	Total volume	Avg urgency prob.	Ag hostpot score
14	sidewalk, dangerous, hole, damaged, broken, bike, …	244,568	0.0838	0.8402
11	thank, hello, regrettably, … (template artifacts)	144,160	0.0860	0.7538
15	tree, height, boards, cardboard, …	37,047	0.0396	1.0183
17	furniture, chair, board, wooden, …	25,911	0.0302	0.9318
0	operator, forwarded, request, …	14,258	0.0376	0.7280
5	bag, white, blue, uncollected, …	11,538	0.0373	0.7329
8	street, corner, dirty, lighting, …	11,176	0.0803	0.5498
4	waste, construction, bin, …	5,788	0.0467	0.6191

 doi:10.20944/preprints202604.1116.v1

19	non compliant, regulatory, parking, …	2,945	0.0223	0.8391
3	deposit, clandestine, illegal, …	3,231	0.0342	0.6109

The z-score based detector did not flag statistically significant events during the analysis window, indicating that fluctuations in report volume and predicted urgency remained below the chosen anomaly threshold. This suggests that, for the evaluated period, the most relevant operational signals are persistent hotspots and chronic issue themes rather than abrupt shocks.

SHAP analyses for the high-urgency class emphasized terms that plausibly encode safety risks, including dangerous, danger, risk, accident, crossing, and references to vulnerable road users (cyclists, pedestrians), as seen in Table 5. Also, for complaint-oriented reports, salient terms included abandoned, bags, clandestine, dirty, illegal, and recurrent. The resulting term lists align with expected municipal semantics and provide an audit trail that can be reviewed by domain experts (see Figure 3).

Table 5. Explainability audit: highest-weighted terms by class (abridged).

Class	Term	Weight
High urgency (class 2)	Dangerous	7.883
High urgency (class 2)	Danger	6.022
High urgency (class 2)	Risk	5.896
High urgency (class 2)	Accident	5.399
High urgency (class 2)	Crossing	4.306
High urgency (class 2)	Cars	3.734
High urgency (class 2)	Falling	3.500
High urgency (class 2)	Cyclists	2.707
High urgency (class 2)	Bike	2.450
High urgency (class 2)	Pedestrians	2.100
High urgency (class 2)	Damage	1.634
High urgency (class 2)	Holes	1.534
High urgency (class 2)	Pedestrian	1.506
High urgency (class 2)	Marking	1.430
High urgency (class 2)	Weeks	1.479
Complaint (class 1)	Abandoned	5.167
Complaint (class 1)	bags	4.322
Complaint (class 1)	Bag	4.216
Complaint (class 1)	Not	3.684
Complaint (class 1)	Clandestine	3.283
Complaint (class 1)	Dirty	3.115
Complaint (class 1)	Again	3.091
Complaint (class 1)	Deposit	2.931
Complaint (class 1)	deposits	2.925
Complaint (class 1)	depot	2.691
Complaint (class 1)	Always	2.440
Complaint (class 1)	Illegal	2.250
Complaint (class 1)	Garbage	2.222
Complaint (class 1)	Depots	2.146
Complaint (class 1)	non	2.146

 doi:10.20944/preprints202604.1116.v1

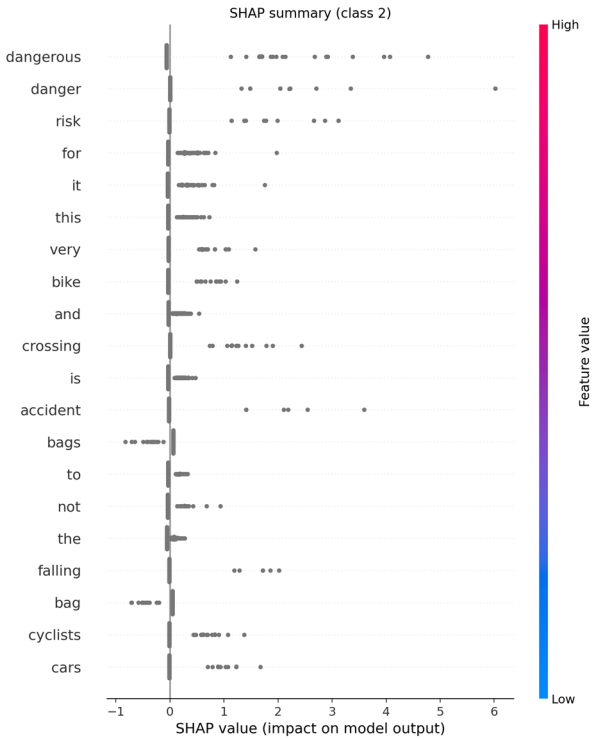
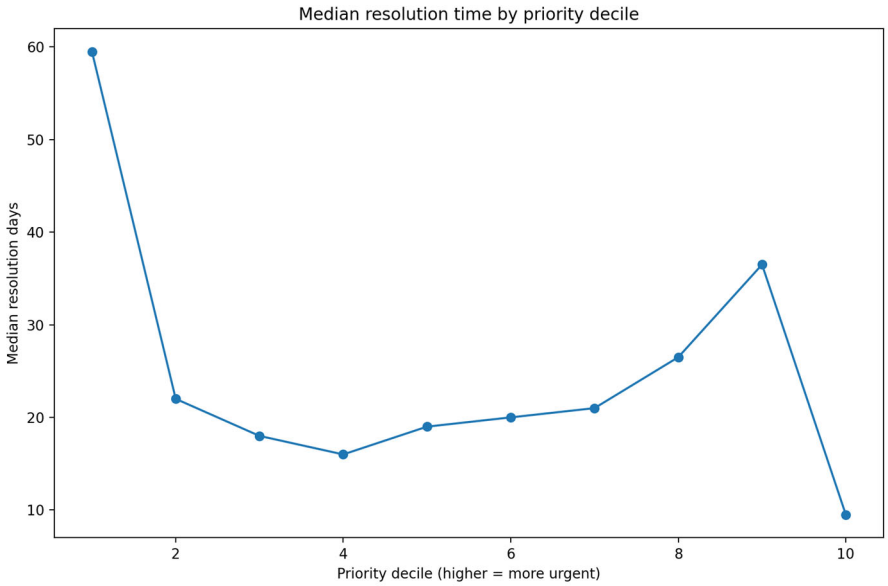


Figure 3. SHAP summary plot for the high-urgency class (class 2) in the TF–IDF logistic regression model. Terms such as dangerous, risk, and accident contribute positively to urgency predictions.

Aggregating closed cases by priority deciles revealed a non-linear relationship between predicted priority and median resolution time. The highest-priority decile showed the shortest median resolution time (approximately 9–10 days), while the lowest-priority decile exhibited the longest (approximately 59–60 days). Middle deciles clustered around 16–22 days, and the second-highest decile showed a longer median (around the mid-30s), consistent with a subset of high-salience but operationally complex cases. A complementary area-week scatter plot showed that very low priority scores coincide with a heavy-tailed distribution of resolution times, including extreme delays, whereas higher priority scores rarely coincide with long resolution times (see Figures 4 and 5).



 doi:10.20944/preprints202604.1116.v1

Figure 4. Median resolution time (days) by priority decile, where higher deciles correspond to higher predicted priority. The top decile is associated with the shortest median resolution time.

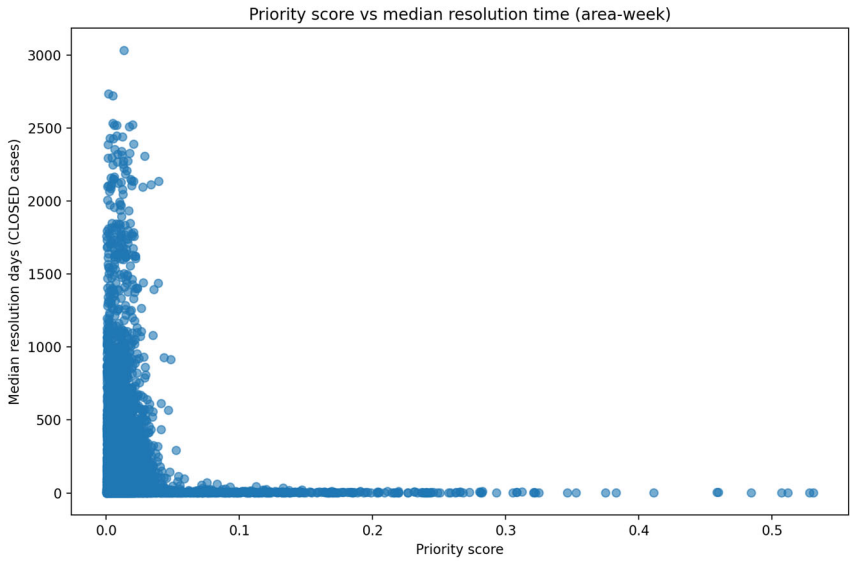


Figure 5. Priority score versus median resolution time aggregated at the area-week level (closed cases). Low priority scores exhibit a heavy-tailed distribution of resolution times.

This pattern is consistent with triage processes that address urgent cases promptly while lower-priority issues accumulate backlog.

4. Discussion

This study set out to translate high-volume, multilingual citizen reports into decision-support signals that remain auditable under municipal constraints. The results suggest that a transparent baseline can already deliver operational value when it is paired with probability calibration and explainability. This is consistent with the broader "citizens as sensors" framing of volunteered geographic information, where citizen input expands monitoring capacity but also introduces heterogeneity and bias that must be handled carefully [1] (pp. 211–221), [2] (pp. 682–703).

A central finding is the trade-off between discrimination and usable uncertainty. The raw TF–IDF + logistic regression model produced strong performance on the held-out gold split, and the confusion matrix indicates near-perfect separation for class 2 in that particular split (10/10 correct). However, the reliability analysis shows the raw model's confidence is not uniformly trustworthy, especially at higher predicted probabilities. In public-service settings where scores may be interpreted as risk signals (not merely ranks), miscalibration can lead to overconfident triage and brittle decision rules. Post-hoc calibration improved probability–accuracy alignment (ECE = 0.069) in the expected direction [10], but it also shifted class decisions toward more conservative predictions, reducing class-2 recall on this split (7/10). Rather than presenting this as a net "win" or "loss," the evidence supports a more practical interpretation: calibration makes the score safer to treat as an estimated probability, while thresholding and workflow design must be tuned to municipal risk tolerance (e.g., prioritizing recall for safety hazards).

The topic and hotspot outputs support a second, complementary use case: moving from individual tickets to recurring themes and persistent spatial pressure. High-volume topics linked to sidewalks/road damage and bulky waste align with the kinds of chronic issues typically surfaced on reporting platforms. The fact that the z-score event detector did not flag shocks suggests that, during the analyzed window, system dynamics were dominated by persistent patterns rather than abrupt anomalies. This is plausible for urban maintenance streams, where seasonality and backlog dominate

 doi:10.20944/preprints202604.1116.v1

variance and where threshold-based detectors intentionally avoid triggering on routine fluctuations. At the same time, the absence of detected events should be interpreted cautiously: aggregation choices (weekly bins, spatial units), the threshold level, and the smoothing implicit in "chronic" workflows can all reduce sensitivity to short-lived disruptions.

Explainability results provide governance-relevant face validity. SHAP attributions identified safety-related lexical cues (dangerous, risk, accident, crossing) and complaint-oriented cues (abandoned, illegal, dirty) as drivers of the corresponding class probabilities, which is consistent with the goal of producing reviewable "hooks" for domain experts [12]. Importantly, explainability also highlighted text hygiene issues: template artifacts (e.g., line-break tokens) appeared in topic terms, illustrating how operational systems can inject boilerplate that contaminates downstream models. This is not merely cosmetic; it can create spurious themes and inflate volume-based signals, which matters when prioritization is tied to hotspot scoring.

The resolution-time analysis suggests an association between predicted priority and operational throughput: high-priority deciles had shorter median resolution times, while low-priority deciles exhibited longer and more variable delays. This pattern is compatible with a functioning triage pipeline, but it should not be interpreted causally. Resolution time is affected by confounding factors such as task complexity, departmental capacity, and administrative routing, and text-derived priority could correlate with these latent variables rather than drive outcomes. The area-week scatter also indicates a heavy-tailed delay regime at low predicted priority, which may reflect backlog accumulation or hard-to-resolve categories rather than systematic neglect. From a decision-support perspective, the main takeaway is that priority scores can be used to structure queues and surface risk pockets, but they should be embedded in a policy-aware workflow (e.g., SLA rules, periodic review of low-priority backlogs) rather than used as an automated dispatcher.

Several limitations bound the conclusions and indicate clear next steps. First, the gold standard is small relative to the corpus, and the high-urgency class is intentionally rare. This is realistic operationally but it increases statistical uncertainty: the confusion matrices shown here contain only 10 class-2 instances in the evaluated split, so recall swings (1.0 vs 0.7) should be treated as fragile estimates rather than stable model properties. Expanding adjudicated high-urgency examples (and documenting inter-annotator agreement) would improve both evaluation reliability and calibration robustness.

Second, English normalization is a pragmatic choice for multilingual modeling, but translation and normalization can erase nuance (especially for short texts) and can introduce systematic artifacts. This issue is visible in the topic vocabulary and in the presence of templated/system phrasing. In future iterations, a bilingual/multilingual approach could be compared against translation-based pipelines using multilingual pretrained models, while keeping interpretability constraints explicit.

Third, the hotspot score and event detector were designed for transparency, but they are not substitutes for formal spatial cluster inference. In particular, volume–urgency aggregation can confound reporting intensity with underlying incidence (e.g., participation biases across neighborhoods), a known concern in volunteered geographic information [2] (pp. 682–703). A careful deployment would therefore combine these signals with municipal context (population, footfall proxies, inspection cycles) and would treat hotspots as triage leads, not ground truth.

Also, platform processes themselves shape the data. FixMyStreet-style systems include administrative routing, status transitions, and templated responses that affect both text and outcomes [13,14]. Model monitoring should explicitly track changes in templates, category taxonomies, and departmental procedures because these can cause silent distribution shifts.

Given the applied, civic-tech focus and the emphasis on transparency and deployment constraints, the work is well aligned with practitioner-facing and "smaller" community venues in digital government and urban analytics. Suitable options include the ACM Digital Government Research community (dg.o) and the EGOV-CeDEM-ePart conference series, as well as specialized workshops collocated with GIS/urban computing events where decision-support prototypes and evaluation on civic datasets are common. Positioning the contribution as an auditable baseline (rather

 doi:10.20944/preprints202604.1116.v1

than claiming state-of-the-art NLP) would match the evidence and better serve the intended audience.

5. Conclusions

This paper presented an explainable, calibration-aware analytics pipeline for multilingual FixMyStreet Brussels reports that integrates urgency modeling, topic discovery, and spatio-temporal hotspot scoring. On a manually annotated gold standard, a TF–IDF logistic regression baseline achieved strong classification performance, and post-hoc calibration improved the trustworthiness of probability estimates at the cost of more conservative high-urgency predictions. Topic-level summaries and hotspot metrics highlighted dominant operational themes (sidewalk/road hazards, bulky waste) and persistent spatial pressure, while thresholder event detection did not identify statistically significant shocks during the analyzed window, suggesting the observed dynamics are primarily chronic rather than episodic. Explainability audits provided linguistically intuitive drivers for urgent and complaint-oriented cases and surfaced data-quality artifacts that matter for governance.

Overall, the results support a cautious but practical conclusion: municipalities can obtain actionable triage structure from simple, auditable models when they treat probabilities as calibrated risk estimates, expose explanations for review, and explicitly account for data-generation biases inherent to citizen reporting. Future work should prioritize expanding the gold standard (especially for rare urgent events), strengthening multilingual handling without increasing opacity, and validating hotspot signals against external operational indicators before using them for resource allocation.

Funding: This research received no external funding.

Data Availability Statement: Data analysed was sourced from https://data.mobility.brussels/en/info/a609a408-4ff1-47df-a0de-d4906bb89469/.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), pp. 211–221.
2. Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design, 37(4), pp. 682–703.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT.
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of ACL.
5. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, pp. 993–1022.
6. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF–IDF procedure. arXiv preprint arXiv:2203.05794.
7. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), pp. 93–115.
8. Kulldorff, M. (1997). A spatial scan statistic. Communications in Statistics—Theory and Methods, 26(6), pp.1481–1496.
9. Truong, C., Oudre, L., & Vayatis, N. (2018). A review of change point detection methods. arXiv preprint arXiv:1801.00718.
10. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of ICML.

 doi:10.20944/preprints202604.1116.v1

11. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of ICML.
12. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems. arXiv:1705.07874.
13. mySociety. (n.d.). FixMyStreet (platform documentation and project resources). Research community references for dissemination venues: ACM dg.o (Digital Government Research). EGOV-CeDEM-ePart conference series.
14. City of Brussels. (n.d.). Report an issue in public space (Fix My Street Brussels). Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. arXiv:1911.02116.