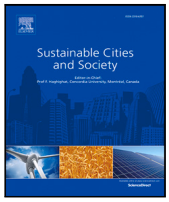
Sustainable Cities and Society 85 (2022) 104026

Contents lists available at ScienceDirect

Sustainable Cities and Society

journal homepage: www.elsevier.com/locate/scs



Machine learning and Natural Language Processing of social media data for event detection in smart cities

Andrei Hodorog *, Ioan Petri, Yacine Rezgui

School of Engineering, Cardiff University, Queen's Buildings, The Parade, Cardiff, CF24 3AA, United Kingdom

ARTICLE INFO

Keywords:
Social media
Smart cities
Event detection
Natural Language Processing
Citizen satisfaction
Machine learning

ABSTRACT

Social media data analysis in a smart city context can represent an efficacious instrument to inform decision making. The manuscript strives to leverage the power of Natural Language Processing (NLP) techniques applied to Twitter messages using supervised learning to achieve real-time automated event detection in smart cities. A semantic-based taxonomy of risks is devised to discover and analyse associated events from data streams, with a view to: (i) read and process, in real-time, published texts (ii) classify each text into one representative real-world category (iii) assign a citizen satisfaction value to each event. To select the language processing models striking the best balance between accuracy and processing speed, we conducted a pre-emptive evaluation, comparing several baseline language models formerly employed by researchers for event classification. A heuristic analysis of several smart cities and community initiatives was conducted, with a view to define real-world scenarios as basis for determining correlations between two or more co-occurring event types and their associated levels of citizen satisfaction, while further considering environmental factors. Based on Multiple Regression Analysis (MRA), we established the relationships between scenario variables, obtaining a variance of 60%–90% between the dependent and independent variables. The selected combination of supervised NLP techniques leverages an accuracy of 88.5%. We found that all regression models had at least one variable below the 0.05 threshold of the $f - test$, therefore at least one statistically significant independent variable. These findings ultimately illustrate how citizens, taking the role of active social sensors, can yield vital data that authorities can use to make educated decisions and sustainably construct smarter cities.

1. Introduction

With social media adoption expanding exponentially and accelerated by the widespread use of mobile devices, users increasingly react in real-time to events occurring in their immediate surroundings and beyond, providing time-critical and potentially actionable information in a smart city context. With population growth and technological advancements, including the deployment of 5G, the potential has become increasingly promising for citizens to act as "active" social sensors and actuators. In 2021, social media was well established, with more than 75% of Americans using at least one platform, according to Faelens et al. (2021). Numerous technical factors contribute to this trend, such as the increasing prevalence of the global Internet availability. According to Russo (2019), in 2018, approximately 99% of European households had access to a 4G connection, compared to 20% in 2012. Another contributing factor is the widespread usage of social media, which governments across the world increasingly employed to communicate directly with citizens (Wirtz, Göttel, Langer, & Thomas, 2020). These breakthroughs enable more and more "always-on" citizens to react in real-time to events, regardless of their location, while openly reporting their internal thought processes to the public.

As social media applications such as Twitter and Facebook surpassed email as the primary communication channel for most users (Lytras, Visvizi, & Jussila, 2020), they accumulated ever-increasing insights of interest to data analysts. One of the leading social media platforms for short-form messages, Twitter witnesses an average daily usage rate of 500 million user posts ("tweets") (Gao et al., 2022). Large proportions of the population of developed countries live in conglomerated urban areas. As described by Malche, Maheshwary, and Kumar (2019), among other functions, smart cities attempt to monitor various parameters such as energy, water quality, and transportation flow while detecting and preventing crime and organising appropriate responses proportionate to identified risks and events.

In an increasingly urbanised and digital world, cities are exploring a wide range of governance models, informed by decision support systems that leverage (near) real-time information, including social media sources, to enhance their sustainability and resilience (Kim & Kim, 2022). As such, the Smart City paradigm seeks to enhance the

* Corresponding author.
E-mail addresses: hodoroga@cardiff.ac.uk, contact@andreihodorog.com (A. Hodorog), petrii@cardiff.ac.uk (I. Petri), rezguiy@cardiff.ac.uk (Y. Rezgui).

https://doi.org/10.1016/j.scs.2022.104026
Received 10 February 2022; Received in revised form 16 May 2022; Accepted 22 June 2022
Available online 2 July 2022

standard of living for its residents. In order to determine whether the development of smart cities is consistent with sustainable development goals, a new methodological approach has gained traction. Girardi and Temporelli (2017) defined *smartainability* as an approach scoped with evaluating, using both qualitative and quantitative measures, how sustainable smart cities become as a result of deploying smart technologies and infrastructure. The author implemented this strategy on the *Expo Milano 2015* site, demonstrating the capability to provide decision-makers with valuable insight into the following advantages generated by the deployment of smart solutions at a city level: (i) benefits can be gauged (ii) indicators can be assessed prior to implementing technologies or solutions (iii) benefits could be easily linked with the associated deployed smart technologies for a more targeted assessment. Our research devises a viable approach to achieving and maintaining *smartainability* by enabling smart city managers, such as stakeholders and local authorities, to make more informed decisions around the detected events, prioritised under the analysis of their associated environmental factors and citizen satisfaction.

1.1. Research motivation and context

Our manuscript proposes a methodology for detecting and interpreting real-time events occurring in a smart city context. A *taxonomy* built upon a semantic-based risk model is presented in this section. We also propose a method to conduct citizen satisfaction analysis, which would enable smart city managers to gauge better the well-being of residents and transients in their areas of interest while also evaluating the importance of particular detected events more objectively. A citizen satisfaction metric is determined based on the *positive* or *negative* valence of each message. In addition, we uncover and validate relationships between co-occurring events, citizen satisfaction, and environmental factors. Furthermore, we validate the relationships using MRA statistical models.

1.1.1. Smart city event detection on a bespoke risk taxonomy

Each message features valuable meta-data, such as originating location and other user-defined attributes and content (images and videos). As a result, manually processing and decoding voluminous data streams becomes increasingly difficult, if not impossible. Notwithstanding this barrier, we can gear the Twitter limitations to our advantage: the highly condensed and standardised format of the messages imposed by the platform, limited to 280 characters each, facilitates the deployment of automated models for aggregating, processing and decoding "events" of interest within the data, with relatively low computational costs.

From a technical perspective, an event represents a real-time occurrence identified within a message that, based on the determination of machine learning NLP techniques, features a semantic similarity above a certain threshold defined in the hyper-parameters. Conversely, from an ontological perspective, an event acts as a stimulus and has the consequence of disturbing the modus operandi of a city, taking place either in a tangible form (*thefts*, *car accidents*, *social gatherings*) or intangible form (*positive* or *negative* sentiments).

An efficient approach to undertaking the risk concerns in a smart city is to react promptly to emerging threats. Utilising the work of Coburn et al. (2014) as a starting point, we have derived a general risk *taxonomy* illustrated in Fig. 1. Events warranting a deeper level of analysis, such as wars, nuclear, space and economic threats, although mentioned and modelled by the original author, have been omitted, as they do not form part of the remit of this research. Although not exhaustive, the *taxonomy* provides a strong starting point for event detection in a smart city context. At this stage of our work, we considered seven categories of events (highlighted within the blue circles) as a starting point. In the taxonomy highlighted above, we have also proposed appropriate responses to each type of event.

1.1.2. Citizen satisfaction and environmental factor analysis

As some decisions taken in a smart city context could directly and significantly affect the livelihoods of the citizens, they should be subject to a periodic review process. Using messages from a social media channel such as Twitter affords us an insight into the publicly exposed thought processes of the residents and transients of a particular area. By analysing *positive* or *negative* inclinations, specific speech patterns and words acquire a *negative* or *positive* value, aiding in the prediction. Inspired by Lamba and Madhusudhan (2018), an extensive collection of tweets was fetched and utilised to ensure a highly accurate category prediction. The extensive vocabulary variations enabled us to account for the subtleties of language and jargon expressing human sentiments.

To analyse environmental factors, we use daily median values for *temperature*, *humidity* and *precipitation* fetched from *Meteoblue*. This platform compiles data from various local weather stations and national weather services and incorporates it into model simulations to provide high precision data at high spatial resolution.

1.1.3. A case study for a Smart City

Amongst several strategic research initiatives, the smart city and communities programmes funded by the European Commission are of particular significance: Smarter Together (Morishita-Steffen et al., 2021), Making city (Gabaldón Moreno, Vélez, Alpagut, Hernández, & Sanz Montalvillo, 2021), CityxChange (Temeljotov Salaj & Loewen, 2020), ATELIER (Baculakova et al., 2020; Castillo-Calzadilla et al., 2021; Olivadese et al., 2021), MAtchUP (Croci & Molteni, 2021; Dimeski, Memeti, & Bogdanoska-Jovanoska, 2019).

A heuristic analysis of these projects, coupled with the recurring event types and use cases modelled as part of our *Taxonomy*, shortlisted several relationships and scenarios of particular significance to smart cities:

- The interplay between *traffic events* and *environmental factors*
- The influence of *faulty infrastructures*, such as *electricity charges*, on *traffic events*
- Relationships between *thefts*, *social events* and *environmental factors*
- Links between *citizen satisfaction* and *gatherings*, coupled with *queues* and *electricity charges*
- Correlations between citizen satisfaction and weather elements (*temperature*, *precipitation*, and *humidity*)

These scenarios resemble real-world use cases in a smart city context, in line with recurrent focal topics highlighted within the European Commission-funded projects mentioned above. We have associated each scenario with a dependent variable and multiple independent variables, all derived from the risk types modelled in the *Taxonomy*. These variables are fully quantified and evaluated statistically in Section 4.

Cardiff City Council has an ambitious agenda for the adoption and implementation of smart city technologies through their released "Smart City Roadmap" (Cardiff City Council, 2020), which outlines proposals for introducing smart street lighting, smart parking, smart transportation, and a smart environment. Most of these initiatives aim at addressing some of the challenges identified: increased demand for public services and energy, uncertain economic conditions, and increased pressure on the natural environment. Our tailored smart city research can assist Cardiff Council and other local authorities in developing an Open-Source Intelligence (OSINT) framework to make evidence-based decisions considering historical and real-time events. This approach enables Cardiff City to maintain a competitive edge in the smart city revolution currently occurring in the UK and to remain a secure and highly available hub for sustainable development.

In a real-world scenario, a smart city manager would monitor the "urban pulse" using an interactive dashboard assessing different types of events at a city level. The events are detected from a continuous stream of messages and classified into a risk category. The events are geo-localised wherever possible and can be validated through manual

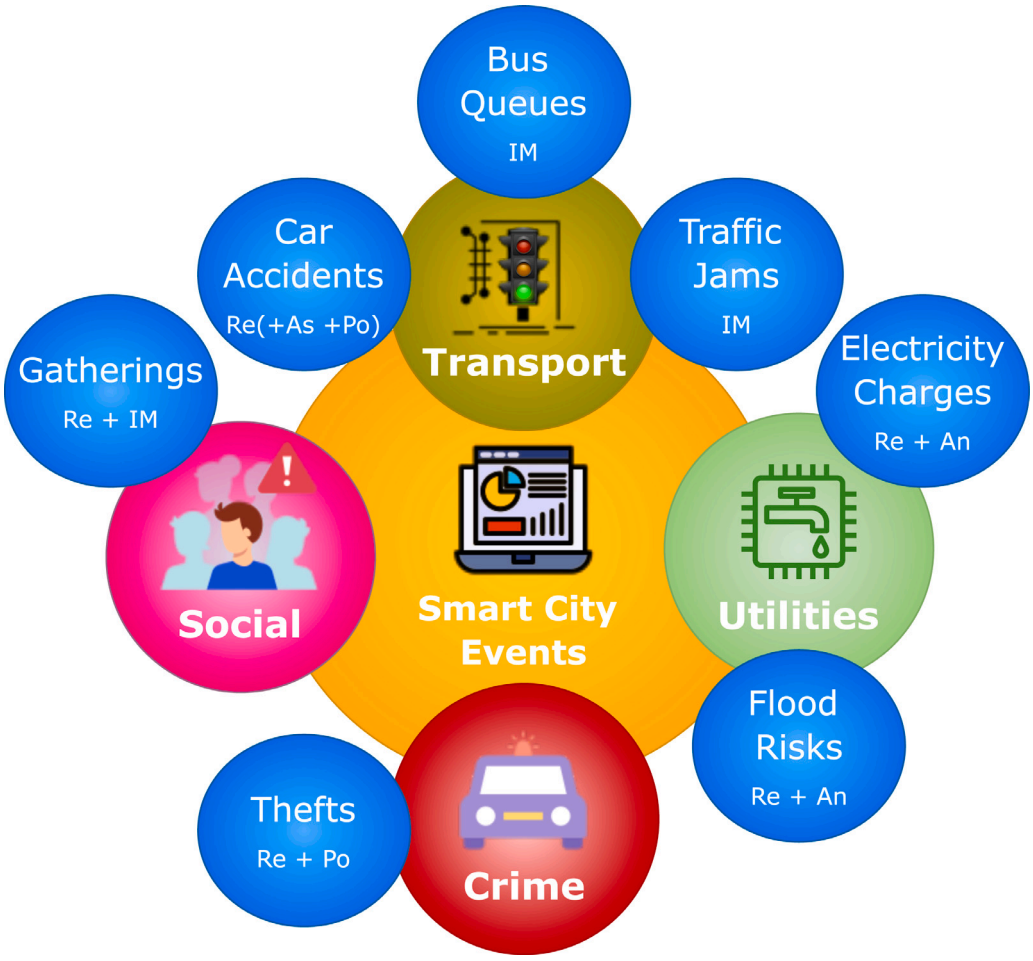


Fig. 1. Smart City Taxonomy example and appropriate local authority responses: IM = Interactive Map, Re = Report, As = Assistance, An = Analyse, Po = Police, Fi = Firefighter, Te = Technician. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

review through cross-checking with third-party data sources, such as intelligence reports from local authorities and public service providers. When significant correlations with citizen satisfaction or environmental factors in the areas of interest are detected, they are displayed automatically alongside the detected events, with suggestions on the significance of the associated event in terms of severity and urgency. The city manager would then conduct a risk assessment and assign proportionate resources, according to an appropriate intervention plan, as per the information gathered in Table 1:

1.1.4. Body of knowledge contributions

As we can observe from the narrative illustrated above, our manuscript covers the following areas of interest: (a) the impact of the social factor on the resilience of cities (b) monitoring and analysing actions and behaviours of citizens within urban communities (c) smart city governance (d) smart city and societies management optimisation through decision support systems for trade-off and uncertainty analysis (e) case studies for Big Data, machine learning, and AI (Artificial Intelligence). We are advancing the body of knowledge in the field of smart cities and sustainable communities by giving context, corroborated by primary sources of evidence, to a wide range of socio-technical and environmental events taking place in an urban landscape through the following contributions:

1. We empower citizens as active sensors in a smart city context, scoped with reporting risks devised in a custom risk taxonomy.
2. We leverage NLP (Natural Language Processing) techniques to make sense of the abundant stream of social media data and detect incidents reported through eyewitness accounts.
3. We determine patterns and trends in risk events occurrences, citizen satisfaction, and environmental factors, at a city level.
4. We analyse relationships between citizen satisfaction, environmental factors and risk events and statistically validate these relationships using Multiple Regression Analysis.

1.2. Research questions and objectives

Our hypothesis is that social media data mining, underpinned by a semantic understanding of the context and conveyed through a taxonomy, provides the means of making informed decisions while promoting citizens to be “active” as opposed to “passive” agents in a smart city landscape. This hypothesis translates into the following research questions:

1. How can we leverage Natural Language Processing (NLP) techniques to make sense of the abundant stream of social media data in a smart city context?
2. What are the general patterns and trends in citizen satisfaction occurring in a smart city context?
3. What are the emerging relationships between these trends and patterns in citizen satisfaction and co-occurring environmental factors and events?

Table 1
Detected risk events and appropriate actions suggested on a dashboard to a Smart City manager.

Event and area	Suggested Action(s)	Priority	Background Information
Fire, Roath park	Dispatch a fire department unit to neutralise the fire	HIGH	• local residents indicated *negative* emotional predilection of *high intensity* • environmental factors (*high temperature*, *low humidity*, *low precipitation*) indicated a *high risk*
Bus queue, Cardiff City Centre	Re-route crowded bus services to increase traffic fluidity	MEDIUM	• local residents indicated *negative* emotional predilection of *moderate* intensity • environmental factors (*low humidity* and *low precipitation*) indicated a *below average risk* to road safety in the short term • no recent *car accidents* or *traffic jams* detected in the area
Flood risk, Cardiff Bay	Actuate the barrage movements to release water into the Celtic Sea	LOW	• local residents indicated *negative* emotional predilection of *low intensity* • environmental factors (*high temperature* and *low precipitation*) indicated *low risk* in the short term

These research questions translate into the following key research objectives:

1. The primary objective of our research is to detect real-time events in a smart city context and determine relationships and dependencies between co-occurring events (such as thefts, car accidents, faulty lights and electricity charges).
2. The second objective of our research is to understand the trend in citizen satisfaction in a smart city context, measured as *positive* or *negative* emotional predilection in the content of their messages.
3. Our third research objective is to understand the nature and extent of the impact of environmental factors and other potentially co-occurring events on citizen satisfaction in a smart city context.

We aim to achieve these objectives through our work exposed in the following sections of our manuscript. Section 2 presents, in addition to the most relevant works identified in the reviewed literature, the edges that our solution brings beyond the State-of-the-Art, outlined side-by-side with their practical and theoretical implications in the form of a summary table.

Section 3 presents the underpinning technical implementation for our experiments: (i) an optimised data pipeline, designed for fast collection and processing, employing parallel computing and fault tolerance; (ii) our choice of data sourcing and classification methods, based upon the results of a pre-emptive evaluation of the accuracy of *AWD-LSTM* combined with *ULMFiT* in comparison with other NLP techniques; (iii) the selection of MRA steps undertaken to validate the models statistically. Section 4 presents the results of our dataset processing and a statistical significance evaluation conducted by applying MRA to the number of occurrences of risk events in correlation with the two associated datasets: citizen satisfaction and environmental factors metrics.

Section 5 highlights our achievements alongside the identified limitations, each accompanied by a reflection on directions for improvement, mitigation and direction for further research. Section 6 presents our conclusions and closing remarks.

2. Related works

When investigating the implications of social media on smart cities, it is vital to understand the stages some pioneering cities have already passed in their deployment (Du et al., 2020). In a context where mobile technology streamlines governments and local authorities, stakeholders have increasingly presented an interest in various KPIs (Key Performance Indicators) at a city level, using social media to facilitate interactions with citizens. According to the findings of Rahimi-Golkhandan, Aslani, and Mohebbi (2021), various applications could assist in understanding the general public emotional predilection, a sensitive factor influencing decision-makers at multiple government levels.

Social media platforms proved effective in identifying the granular specifics of local communities, playing a pivotal role in promoting accountability and transparency within society (Bellini et al., 2021). A significant amount of evidence pointed out that UK government officials utilised social media platforms for accountability in governing, which also assisted stakeholders and policymakers in a more productive and accurate analysis of a series of events that had previously seemed unrelated (Yuan et al., 2020). As confirmed by Sharida, Hamdan, and Mukhtar (2020), social media led to enhanced levels of engagement and deeper trust with the public in the context of smart cities. Hodorog, Petri, Rezgui, and Hippolyte (2021) demonstrated, through the use of IDF and Metric-Cluster techniques, that multi-disciplinary collaboration on social media platforms between white and blue-collar workers in the construction industry sector led to the identification of training gaps.

In terms of the ideology and conceptualisation of smart cities, the current generation of Internet users tended to focus on the positive aspects of social media, establishing a well-functioning government strategy that promotes cooperation and collaboration among citizens and external organisations (Rehman et al., 2020). This approach decreased the number of government-sponsored projects whilst increasing the level of responsibility held by residents at the same time. Kummitha and Crutzen (2019) concluded that a city's institutional framework could be enhanced by encouraging citizen participation and encouraging involvement in the decision-making processes. Therefore, it is vital to ensure that smart city citizens are adequately informed, and mutual trust is established between citizens and government institutions. Bellini et al. (2021) stated that innovative governance and digital mass media stability should be achieved in all smart cities, especially where political management represents a highly sensitive issue, as the emerging social tensions could pose a significant threat to the development and sustainability of a smart city (Colding, Colding, & Barthel, 2020).

Salminen et al. (2020) recently developed an innovative solution for detecting instances of weak trust in authorities and conflicts between citizens. The author experimented with several classification models detecting hate speech on multiple social media platforms and created a holistic model that outperforms the previous keyword-based baseline classifiers. With the spread of COVID-19, Yang, Xiu, Sun, Ying, and Muthu (2022) attempted to determine the intricacies of misinformation spread through social media. Data from Sina Weibo, one of China's most popular social media networks, was used to test theoretical models, revealing that seeking health advice and emotional support intensifies the risk of spreading misinformation. Ancillary content posted on social media platforms, such as images and videos, provides an invaluable insight into citizens' emotional predilection and behaviour in a smart city context. Zhao et al. (2019) attempted to use multi-modal sentiment analysis by extracting and analysing the visual features of ancillary content of social media posts, proposing an additional semantic feature for sentiment classification.

Event detection has attracted extensive research when applied on social media and in other smart city infrastructure contexts, such as energy optimisation. Harnessing on the plethora of valuable insights delivered by semantics, Li, Rezgui, and Kubicki (2020) devised an

Table 2
Gaps identified in the reviewed literature vs. our progress beyond State-of-the-Art.

Reviewed publication	Proposed method of the related work	Gaps/Our progress beyond State-of-the-Art
Salminen et al. (2020)	Detecting instances of weak trust in authorities and conflicts between citizens using several holistic classification models identifying hate speech.	Our proposed AWD-LSTM combination is more efficient than keyword-based baseline classifiers, as it can identify, analyse and contextualise events faster with more accuracy.
Yang et al. (2022)	Evaluating misinformation spread through social media. The model uses conventional computer processing, which cannot cope with real-time analysis.	Conventional text processing unsustainable for real-time analysis. Our combination of NLP techniques leverages parallel computation methods, which render them suitable for real-time event detection.
Zhao et al. (2019)	Analysing citizen satisfaction through analysis of multimedia content posted by users on their profiles through image-text classifiers using machine learning models.	Multimedia content involves resource-intensive computational tasks that are not always available. Our approach is not dependent on the presence of images or videos.
Li et al. (2020)	Real-time response management of a thermal grid for energy optimisation and CO2 emission reduction using text-based event detection.	Text analysis in the absence of NLP or machine learning is a static approach which is not adaptive to new contexts. We utilise supervised learning techniques, which are highly dynamic and tailored to the particulars of the dataset artefacts, while leveraging semantics for the contextualisation of events in a smart city context.

approach for real-time response management of a thermal grid. An intelligent semantic network provided holistic characterisation before optimising the operational schedules of heat generation results. The authors considered a wide range of factors at a building and district level, continuously calibrating models that reflect real-world operational scenarios in a smart city context. The continuous collection and integration of real-time data from heterogeneous sources, in conjunction with a comprehensive feedback loop based on semantics, led to a 36% reduction in operational costs and an alleviation of 43% in CO_2 emissions.

As we can observe, the literature on social media data monitoring is abundant. Previous works focus on various use cases for event detection, such as detecting citizen dissatisfaction with local authorities, highlighting hate speech, evaluating misinformation, and optimising energy efficiency. While some works utilise machine learning, the paradigms adopted, particularly in the absence of parallel computational approaches, are not as sustainable for real-time event detection as our combination of proposed NLP techniques. Moreover, few of the existing related manuscripts focus on a city level, and even fewer make use of a bespoke risk taxonomy to categorise (nearly) real-time events. We devised a summary comparison in Table 2 to reflect on these observations:

As we can follow from Table 2, our manuscript proposes a novel technique for conducting monitoring, analysis and change within urban communities, including analysis for improved management of cities using machine learning and NLP applied to the dataset and requirements of a real-world smart city case study. The following section describes this approach in detail.

3. Methodological approach

With a clear delimitation of the scope of our analysis defined in Section 1.1, we could then outline our approach to data sourcing, constructing the source dataset and pre-processing. This section presents a high-level overview of our application architecture (Section 3.1), together with our selected language classification models, based on a comparative pre-emptive evaluation, and a baseline definition for event detection (Section 3.2).

To meet the research objectives specified in Section 1.2, we adopt a multi-stage methodology, as presented in Fig. 2. The process starts with a data pipeline based on raw data. Tweets are fetched from the Cardiff City area and filtered as closely as possible within the geographical delimitation. The sentences of the messages retrieved are then pre-processed and filtered by keywords and *cosine similarity*. The text is then sanitised (curated) using regular expressions and assigning categories with the aid of the *Classifier*. The labelled dataset, put in conjunction with the useful insight gathered by regular expressions, is now modelled and ready to generate responses, formulating the historical dataset and splitting it into areas (neighbourhoods) and time intervals. Finally, the categorised dataset is modelled using MRA for correlations between co-occurring events, citizen satisfaction (*positive/negative* texts) and environmental factors (*temperature*, *precipitation*, and *humidity*).

3.1. Data pipeline and application architecture

Fig. 3 presents a high-level architectural overview of our technical setup. We designed a redundant data pipeline to expedite the retrieval process, spreading the workload over multiple machines.

To abstract the underlying OS (Operating System) for uniformity and resource compartmentalisation, a virtual network of *Docker* containers (Cito et al., 2017) was deployed using *Debian*-based Linux image environments. Users can query data through the *API (Application Programming Interface)*, specifying the date interval, area and other parameters. This architectural implementation, coupled with a *Message Broker* conducting parallel processing aided by the RabbitMQ library (Ionescu, 2015), enabled us to afford the collection and classification of tweets in (nearly) real-time.

We executed ten parallel classifier instances as part of a High-Performance Computing (HPC) cluster of machines. The delegation of processing tasks takes into account: (a) the length of the message needing to be processed (b) the semantic complexity of the keywords used as part of the tweet message (c) the presence of polluting factors, such as stop words, punctuation, orthography, quotes from movies, music lyrics, para-verbal written language (emoticons/smileys).

The *Tweet Retriever Coordinator* is a simple automated Python script listening to the queue of published intervals while simultaneously connecting to Twitter and downloading tweets matching specific criteria (most commonly, time intervals, keywords and geo-coordinates). This process is scheduled to publish all intervals for the required historical period to the *Message Broker*.

After retrieval, each tweet is published as an individual message to the *Message Broker* to be processed. The *Message Broker* is the central point of the application, enabling swiftly synchronisation between processes and message buffering while also avoiding workload duplication. Workers would request the processing of new messages at their own pace in a non-symmetrical usage model, redistributing their load into multiple segments running on different machines when their I/O (Input/Output) usage passes a certain threshold. Upon completing an interval, the resulting tweets would be en-queued as individual messages to be processed.

This language model is trained to predict the next word based on the previous words taken as input. During this phase, the neural network understands the English language and vocabulary. The *category* model is then trained on the dataset, as per the methods described in Section 3.2. Once the language models are trained on the dataset, we

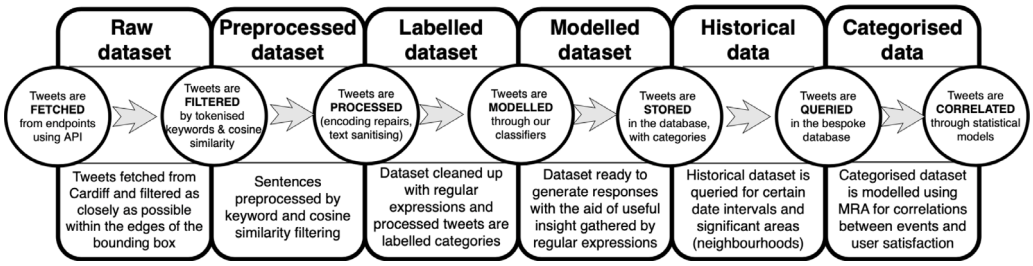


Fig. 2. A high-level overview of the steps implemented as part of our proposed approach.

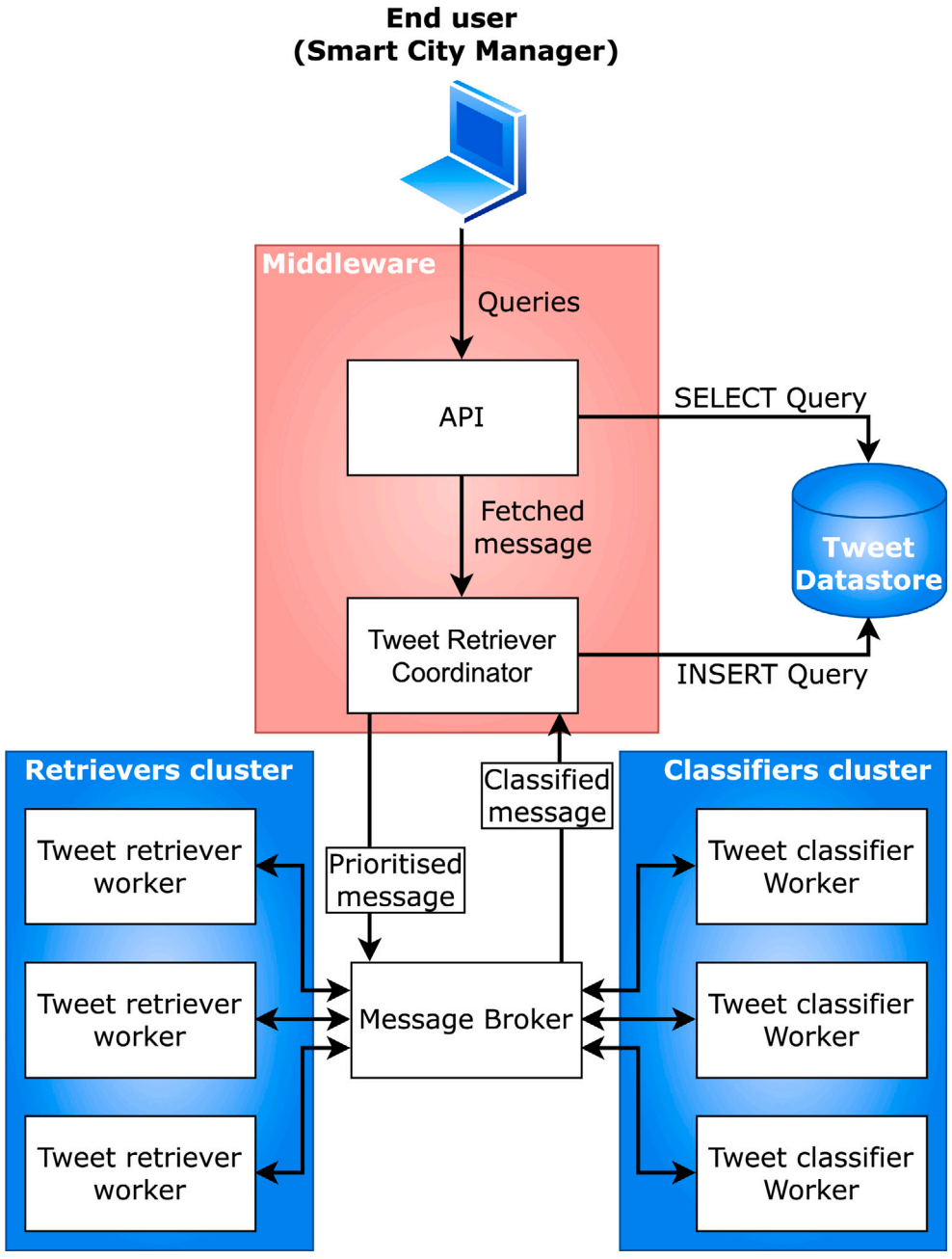


Fig. 3. Data pipeline within the three modules of our application (*Middle-ware*, *Retrievers cluster*, *Classifiers cluster*). The *Tweet Retriever Coordinator* splits the messaging interval into smaller chunks and publishes them to the *Message Broker*.

train the *Classifiers*, comprising linear layers placed at the end of the encoder of the language model. Each classifier instance loads the two models (*category* and *emotional predilection*), then requests messages to be processed and determines, for each of them, the label and sentiment.

To store the models and pre-load them, we used the *Pickle* serialisation library (Fasnacht, 2018). Upon receiving a message, the text is pre-processed, classified, and finally published back to the *Message Broker*. To avoid encumbering the *Classifier*, we stored the curated tweets in a separate database, along with the messages retrieved from the *Message Broker*. The asynchronous modus operandi enabled the *Classifier* to work at total capacity despite the I/O (Input/Output) load.

3.2. Dataset construction and processing

As Kalinin, Krundyshev, and Zegzhda (2021) mentioned, smart cities are about preventing risks besides analysing the quality of services and procedures that differ from one city to another. Therefore, we carefully selected a classification with a spectrum of applicability as broad as possible. This approach enables us to demonstrate the viability of our technical solution beyond the confines of the case study illustrated in Section 1.1.3. For data filtering and classification. Our *Classifier* inherits the same seven event types modelled as part of this *Taxonomy*.

3.2.1. Processing models and pre-emptive testing

Before embarking on the collection and processing of tweets, we considered a baseline model for event detection. We preemptively evaluated several language classification models identified in the literature review. For baselines, tweets are converted into a matrix of token counts before considering the following three types of classifiers (language models used for classifying texts):

- *MNB (Multinomial Naive Bayes Classifier)*, computing the probabilities of belonging to a class as a function of the occurrence of different words (featuring high popularity because of its simplicity).
- *CNB (Complement Naive Bayes Classifier)*, utilising the same principle behind MNB while correcting its assumptions and rendering it suitable for imbalanced data (better suited for conversational tweets rather than informative tweets).
- *RF (Random Forest Classifier)*, providing accurate baselines on regression and classification tasks

We tested our *Classifiers*, using the *Scikit Learn* machine learning library (Kramer, 2016), with the hyper-parameters defined as follows: $alpha = 0.5$ for *MNB*; $alpha = 1.5$ for *CNB*; $n\_estimators = 50$ for *RF*. The primary comparison metric is the accuracy, measured over *five folds*. As expected, the CNB classifier outperforms the MNB. Models which incorporate a pre-processing stage led to more accurate predictions than their counterparts processing raw data. Some complex hashtags are altered and favour the model with no pre-processing because it can label tweets containing the same hashtag. However, we observed this gap narrowed as we collected additional data, reducing the bias induced by the classifier to the maximum extent possible.

As a classification model for the data mining and classification stages we use UMLFiT (Universal Language Model Fine-Tuning), a technique created by devised by Howard and Ruder (2018), which pre-trains a LM (Language Model) on a sizeable general-domain corpus and then uses novel techniques to fine-tune it for the target task. The method is universal in the sense that it meets the following practical requirements: (i) it applies to tasks with varying document sizes, numbers, and types of labels; (ii) it is built on a single architecture and training process; (iii) no custom feature engineering or pre-processing is required; and (iv) no additional in-domain documents or labels are required. ULMFiT comprises the following steps:

Table 3
Accuracy of the previously utilised classification models vs. our combination (*AWD-LSTM* and *ULMFiT*).

Model type	Without tweet pre-processing	With tweet pre-processing
MNB	78.2% (± 3.4%)	76.6% (± 4.6%)
CNB	80.2% (± 2.7%)	79.6% (± 3.6%)
RF	85.0% (± 5.4%)	83.2% (± 6.6%)
AWD-LSTM & ULMFiT	**88.5**% (± **3.2**%)	**88.4**% (± **2.0**%)

Table 4
The filtering Cardiff City bounding box coordinates (TLP = Top-Left Point, BRP = Bottom-Right Point).

TLP longitude	TLP latitude	BRP longitude	BRP latitude
−3.325587	51.444454	−3.067674	51.554148

1. *General-Domain LM Pretraining*: the LM is pre-trained on a large general-domain corpus to forecast the following word in a sequence (with a particular degree of certainty). At this stage, the model acquires knowledge of general linguistic characteristics, such as the typical sentence structure of "subject-verb-object" of an English language sentence.
2. *Target Task LM Fine-Tuning*: the LM is fine-tuned on the data of the target task, acquiring task-specific features of the language elements (in our case, Twitter messages), such as the existence of handles, the usage of slang, abbreviated words, and emojis.
3. *Target Task Classifier*: as a third step, the pre-trained LM is expanded by two linear blocks so that the final output is a probability distribution over the sentiment labels (i.e. *positive* or *negative*), and a label from the risk taxonomy, respectively.

For the pre-training steps of ULMFiT, we propose the innovative language model AWD-LSTM (Merity, Keskar, & Socher, 2017), which encompasses a standard LSTM (Long Short-Term Memory language model) with various tuned dropout hyper-parameters (which involve no additional complexity layers, such as attention or shortcut connections). ULMFiT enables the adaptation of a pre-trained model, fine-tuning the neural network layers to tailor it to our research objectives. ULMFiT is integrated with our dataset through the *Python*-based *fast.ai* library (Howard & Gugger, 2020), which facilitates the freezing/unfreezing of layers, offering a fine-tuned approach to customising the model.

As we can observe in Table 3, this combination of techniques has exhibited a higher accuracy than its counterparts based on the measured parameters. While acknowledging the susceptibility to over-fitting of this approach, we have mitigated this by employing a *stratified K-fold cross-validation*. More insight into the technical setup underpinning the association of these two NLP techniques can be found in a recent article by Briskilal and Subalalitha (2022).

3.2.2. Dataset scope, sourcing and filtering

Our target dataset consisted of messages broadcast by people residing in or transitioning through the City of Cardiff, United Kingdom. For the purpose of our manuscript, we only analysed tweets written in English.

A bounding box for geo-fencing (Table 4) was defined for filtering live data stream, such that only tweets from Cardiff were collected.

The *Twint* Python library (Xavier & Souza, 2020) was used to retrieve the tweets, enabling timely collection without utilising the Twitter API (Application Programming Interface), which could have been subject to call limitations. For data anonymisation purposes, once a message is fetched, only a generic unique *id* of the tweet, *date* and *text* was stored in our database to avoid any links that could be established with the identities of actual persons.

While a significant proportion of the collected tweets were not geo-located (because of users not having the "location sharing" function

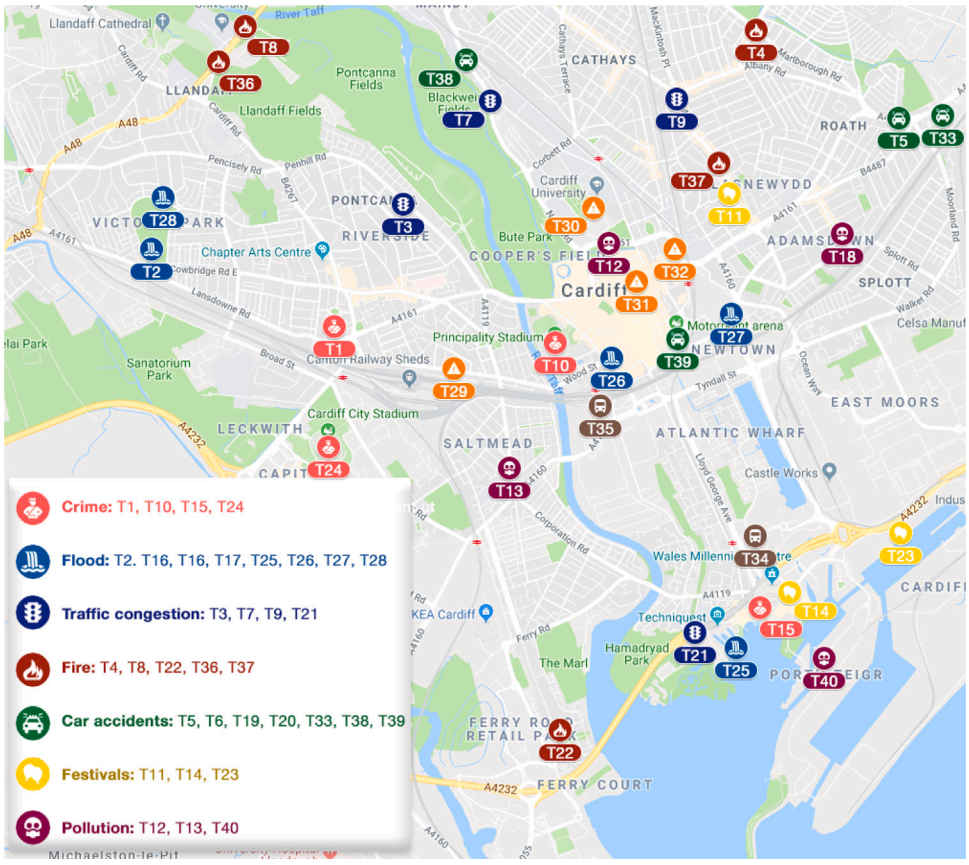


Fig. 4. The interactive map of events in the city of Cardiff derived from the initial results produced by our classifiers.

enabled on their devices), we determined the users' location through other means wherever possible. Another mini neural network was trained to predict originating neighbourhoods or landmarks based on the *biography* of the user and additional information shared within their profile (connections, hashtags and other references to locations in their tweets).

An initial test data subset generated the citizen-accessible interactive map illustrated in Fig. 4. Following our initial promising results, we proceeded to deploy our production environment for collecting tweets. 1.6 million tweets over a period of nine months in 2020.

3.2.3. Dataset pre-processing

Before feeding the contents of the tweets into our model, we pre-processed them to make them ready to be parsed by the tokeniser. This process represents an additional safeguard to ensure that as much information is retained from the original content as possible, while also avoiding classification bias occurring within the model (for instance, associating an event to a specific date). The tweets were filtered using keywords and hashtags, adjusting the generality of the filtering rules to prevent inducing bias in the training process of the *Classifier*. Tweets are saved within the data-store, alongside their semantic selection criteria (the keywords and/or hashtags used for identification). To account for lexical diversity, similar tweets were further filtered using the *cosine similarity* approach (Xia, Zhang, & Li, 2015).

For the text of each tweet: (i) encoding errors are repaired (ii) retweets, hyperlinks, emails, phone numbers, "hashtags", "cashtags", signs, date, time, smileys and emojis are replaced by unique tokens (iii) text written in "CamelCase" syntax is split, and unique tokens are added at the start of the corresponding expressions (iv) unnecessary white spaces are removed. To increase the quality of this process, we use carefully crafted regular expressions, which minimise the number of false positives. Before inputting the text in the model, the final step is assigning tokens using the *fast.ai* tokeniser (which also provides access to the ULMFiT language model).

3.3. Multiple regression analysis applied on the trends identified as part of the scenarios

Being a supervised learning task by itself, we considered MRA a highly suitable and accessible statistical instrument to validate the preliminary correlations determined by our supervised learning NLP techniques. Given the predominantly linear nature of the relationships between the variables forming our scenarios, we considered MRA to strike a favourable compromise between complexity and accuracy. The results are easier to interpret in contrast with the output of other algorithms. The stages of the MRA validation were tailored to our research objectives defined in Section 1.2 and independently applied to each scenario.

3.3.1. General regression statistics and ANOVA

The following general regression statistics are applied to all our scenarios: *Multiple Determination Ratio R*, R^2, *adjusted* R^2 $= (\sqrt{\frac{ESS}{TSS}})$, and the *Standard Error (SE)*. *Multiple R* indicates the intensity and nature of the connection between the variables. The *Adjusted* R^2 indicates the percentage of the variation of the independent variable that can be explained by the simultaneous variation of the independent variables. Unlike the R^2, the *Adjusted* R^2 considers the degrees of freedom in addition to the number of parameters included in the model.

Standard Error (SE) is the approximate standard deviation of a statistical sample population, which measures the accuracy of a sample distribution representation of a sample population. A "sample mean" deviates from the actual mean of a population sample, and this deviation is the standard error of the mean. The *SE* can also be explained as the difference between the expected value and the actual value of the

variables. All our scenarios exhibited a relatively small *SE*, which is an indication that the sample mean is an accurate reflection of the actual population mean.

For each scenario, we analyse the significance of the model using the *Fisher test* applied to the *ANOVA (Analysis of Variance)* table. ANOVA is a statistical analysis tool used to test the degree of differences between two or more groups forming part of an experiment. The results of the ANOVA test are displayed in a tabular form known as an ANOVA table, which displays the statistics used to test hypotheses about the population means (testing it with intra-group and inter-group variants). For each scenario, we are formulating the following two hypotheses: (i) $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$ (the model is not statistically significant) and (ii) $H_1 : \beta_0 \neq 0$ or $\beta_1 \neq 0$ or $\beta_2 \neq 0$ (the model significantly explains the connection between the variables). Based on the value of the *Fisher test (f-test)*, we decide whether to reject the H_0 hypothesis and what is the risk that we apply. If we reject H_0, we conclude that the model is statistically significant (for 5% risk). We then broadly determine the type of dependence between our variables.

3.3.2. Analysis of regression coefficients and variability for significant parameters

The second step for analysing each scenario is applying the *Student's* $t-test$ on the regression coefficients, considering the estimators obtained using the *Least Squares* method and their distribution law. Firstly, we are writing the multiple linear regression model equation using the following formulas: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ and $Y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$. In both formulas x_1, x_2 and x_3 are the values of the *independent variables*, Y is the value of the *dependent variable*. In the first formula, β_0, β_1, β_2, and β_3 are the *hypotheses variables* associated with the significance of each parameter. In the second formula, b_0 is the *coefficient of the dependent variable* and b_1, b_2 and b_3 are the *coefficients of the independent variables*.

For each β parameter above, we follow the classical steps of MRA statistical evaluation:

1. formulating the two hypotheses ($H_0 : \beta = 0$ and $H_1 : \beta \neq 0$)
2. selecting a significance threshold $\alpha = 0.05$
3. choosing the statistical test t and an associated theoretical value of the statistic $t_{\frac{\alpha}{2},n-k}$
4. calculating the test value for the associated coefficients: $\frac{b_1}{s_{\hat{\beta}1}}$
5. for each parameter in the linear regression model equation, we apply the following decision rule: if $\sigma < \alpha$, we reject H_0, and we conclude that the parameter β is statistically significant (for 5% risk)
6. for each parameter identified as "significant", we showcase the correlation between its variability and the variability of the dependent variable, quantifying the variations that occur, on average

3.3.3. Residuals

To further determine the statistical significance of the regression models corresponding to each scenario and further verify their accuracy and relevance, we analyse the following metrics and plots where appropriate: (i) plots of residuals vs. fitted values, (ii) case order plots of leverage (iii) normal probability plot of residuals (iv) histogram plot of residuals (v) residuals vs. observation order.

4. Results and statistical validation

This section presents the results obtained by our proposed data harvesting and classification modules. The scenarios analysed within this section use events previously detected following the data pipeline defined in Section 3.1 and leverage the algorithms and methods defined in Section 3.2. The motivation and reasoning behind the selection of scenarios have been outlined in Section 1.1.1.

At the start of the analysis of each scenario, we outline the general trends in the number of occurrences of risk-based events, on a daily time-frame, present within the evaluated period in year 2020. We firstly infer a subjective correlation between the dependent and independent variables based on these general trends. In the second part of each scenario analysis, we statistically validate the previously inferred subjective correlations by applying MRA statistical models on the classified datasets.

As we can observe from the graphical trend representations outlining occurrences in the sections below, Scenarios 1–5 consistently featured an intermittent reduction in activity between August and September. This pattern could potentially be explained by COVID-related uncertainties and restrictions, co-occurring with the start of a season characterised by lower temperatures and higher precipitation and humidity.

Table 5
Regression statistics for Scenario 1.1.

Multiple R	0.62
R^2	0.39
Adjusted R^2	0.38
Standard error	1.51

Table 6
ANOVA table for Scenario 1.1.

	SS	MS	F	Sig. F
Regression	371.79	123.93	54.05	4.92E–27
Residual	580.03	2.29		
Total	951.82			

4.1. Scenario 1: Dependence of car accidents on congestion and environmental factors

In the first part of the current scenario (Scenario 1.1), we considered *car accidents* as a dependent variable and *congestion*, *temperature* and *humidity* as independent variables. In the second part (Scenario 1.2), we replaced *humidity* with *precipitation* as a dependent variable.

4.1.1. Preliminary results

From Fig. 5, we can infer a general *positive correlation* between *congestion*, *humidity* and *car accidents* throughout the evaluated period, with peaks in activity starting early July and ending early-mid September.

This trend could be explained by the summer period, generally accompanied by more activity on the road due to the summer holiday season. In general, the days with *peaks* in congestion and humidity also feature *peaks* in car accidents. This occurrence repeats for days with *dips* in both metrics. We can infer a general *positive correlation* between *congestion*, *humidity* and *car accidents*.

4.1.2. Regression statistics and ANOVA — Scenario 1.1

In Table 5, $R^2 = 0.39$ suggests a connection of low, but positive intensity between variables. *Adjusted* $R^2 = 0.38$ shows that 38% of the variation of the dependent variable (*car accidents*) can be explained by the simultaneous variation of the independent variables (*congestion*, *temperature* and *humidity*). In Table 6, the small value of the $f-test = 4.92231E-27$ leads us to reject H_0 for both 1% and 5% risk and state that a linear statistical dependence of medium intensity exists between the variables.

4.1.3. Regression coefficients and variability — Scenario 1.1

Table 7 dictates the multiple linear regression model equation: $Y = -2.192 + 0.146 * x_1 + 0.092 * x_2 + 0.23 * x_3$, where $x_1 =$ *congestion*, $x_2 =$ *temperature*, and $x_3 =$ *humidity*.

After computing the *t-tests*, we reject H_0 due to $\sigma < \alpha$. All parameters are statistically significant. For β_0 (*car accidents*), $\frac{b_0}{s_{\hat{\beta}0}} =$

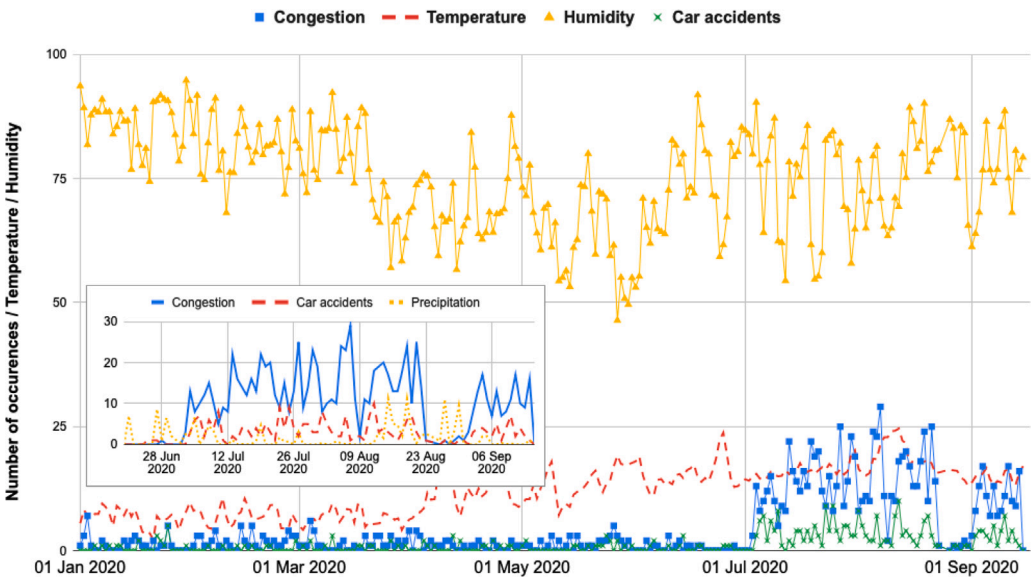


Fig. 5. Preliminary results of social media data analysis for Scenario 1 — Smart City event occurrences.

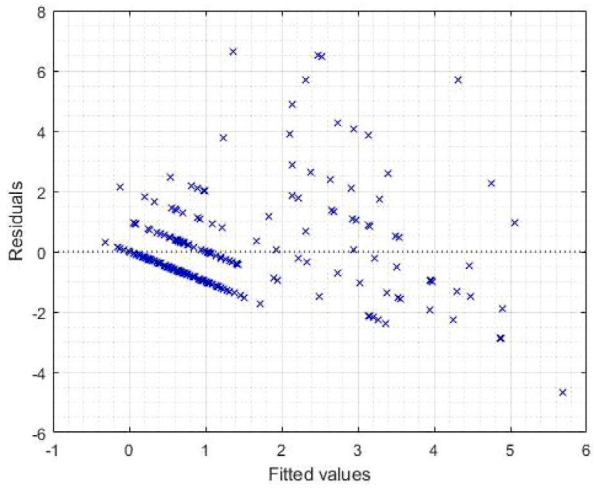


Fig. 6. Plot of residuals vs. fitted values — Scenario 1.1.

$-2.192/0.86 = -2.5$. For β_1 (*congestion*), $\frac{b_1}{s_{\beta 1}} = 0.146/0.017 = 8.6$. For β_2 (*temperature*), $\frac{b_2}{s_{\beta 2}} = 0.092/0.024 = 3.83$. For β_3 (*humidity*), $\frac{b_3}{s_{\beta 3}} = 0.023/0.0097 = 2.37$.

For β_0 (*car accidents*), the average is -2.192 when all other variables are simultaneously equal to 0. For β_1, it increases by 0.1466 when *congestion* increases by *1 unit*. For β_2, it increases by 0.092 when *temperature* increases by 1 degree. For β_3, it increases by 0.023 when *humidity increases* by *1 unit*. The coefficients formula also confirms that the *congestion* variable has the strongest positive effect on the dependent variable (*car accidents*), while *humidity* has the least positive effect.

4.1.4. Residual analysis — Scenario 1.1

The plot of residual distribution over the fitted values in Fig. 6 suggests a clear conical pattern emerging, which could be narrowed to either missing variables in the model or dependencies present in the independent variables (*humidity* could be a function of *temperature*). However, a large proportion of the variation of the dependent variables can still be explained by the selected "independent" variables, suggesting a relatively significant model that warrants improvements.

Table 7
Regression coefficients for Scenario 1.1.

	Coefficients	Standard error	t Stat	P-value
Intercept	−2.192666185	0.8666880374	−2.52993706	0.01201601416
Congestion	0.1466703356	0.01793238305	8.179076658	1.39E−14
Temperature	0.09212658428	0.02417168049	3.811343789	0.000173632911
Humidity	0.02302590679	0.009790444796	2.351875453	0.019446223

Table 8
Regression statistics for Scenario 1.2.

Multiple R	0.68
R^2	0.46
Adjusted R^2	0.45
Standard error	1.41

Table 9
ANOVA table for Scenario 1.2.

	SS	MS	F	Sig. F
Regression	359.19	119.73	51.11	7.34E−26
Residual	592.63	2.34		
Total	951.82			

4.1.5. Regression statistics and ANOVA — Scenario 1.2

This Section commences the second part of our current scenario analysis, with the independent variable interchanges, as described at the start of Section 4.1.

In Table 8, $R^2 = 0.46$ suggests a direct connection between the four variables, of *medium* intensity. *Adjusted* $R^2 = 0.45$ shows that 45% of the variation of *car accidents* variable can be explained by the variations of *faulty lights* and *electricity charges*. The low value of f-test = 7.34212E−26 in Table 9 leads us to reject H_0 for both 1% and 5% risk and state that the model significantly explains the connection between the variables, exhibiting a linear statistical dependence.

4.1.6. Regression coefficients and variability — Scenario 1.2

Table 10 dictates the multiple linear regression model equation: $Y = -0.2623 + 0.1546 * x_1 + 0.0718 * x_2 + 0.0046 * x_3$, where x_1 = *congestion*, x_2 = *temperature*, and x_3 = *precipitation*.

For β_0 (*car accidents*), since $\frac{b_0}{s_{\beta 0}} = 0.3277/0.1095 = 2.992$ and $\sigma > \alpha$, we do not reject H_0 (β_0 is not statistically significant). All other parameters are statistically significant, as we reject H_0 due to $\sigma < \alpha$

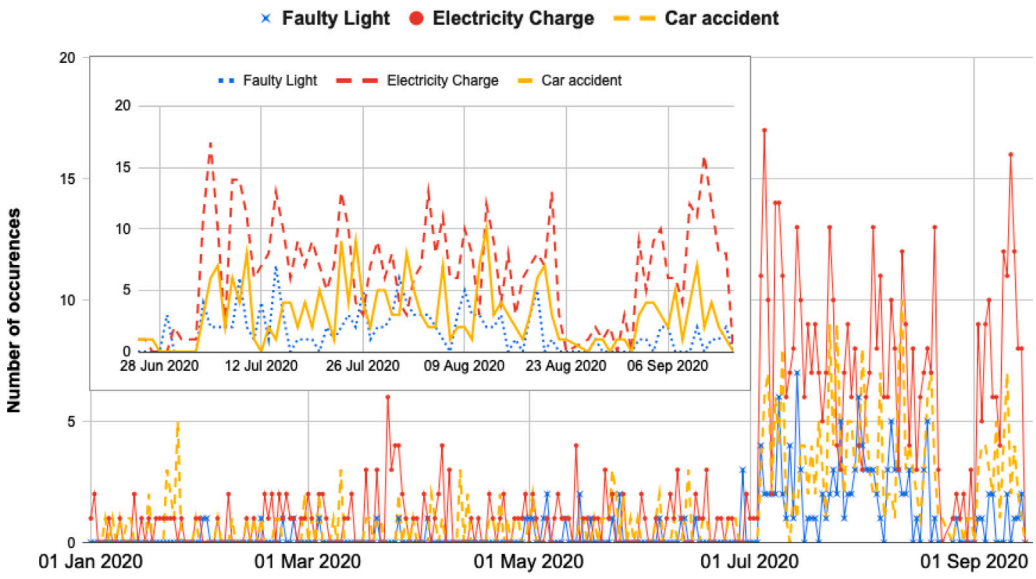


Fig. 7. Preliminary results of social media data analysis for Scenario 2 — Smart City event occurrences.

Table 10
Regression coefficients for Scenario 1.2.

	Coefficients	Standard error	t Stat	P-value
Intercept	−0.2623316889	0.2746086367	−0.9552929289	0.3403414577
Congestion	0.1546656002	0.01779547431	8.691288442	4.57E−16
Temperature	0.07181207595	0.0229204853	3.133095788	0.001933427905
Precipitation	0.004672604321	0.0257001739	0.1818121674	0.8558757174

in the *t-tests*. For β_1 (*congestion*), $\frac{b_1}{s_{\hat{\beta}1}} = 0.3277/0.1095 = 2.992$. For β_2 (*temperature*), $\frac{b_2}{s_{\hat{\beta}2}} = 0.3277/0.1095 = 2.992$. For β_3 (*precipitation*), $\frac{b_3}{s_{\hat{\beta}3}} = 0.0046/0.025 = 0.18$. For β_0 (*car accidents*), the average is −0.26233 when the other variables are simultaneously equal to 0. For β_1, it increases by 0.154 when *congestion* increases by *1 unit*. For β_2, it increases by 0.07 when *temperature* increases by *1 unit*. For β_3, it increases by 0.0046 when *precipitation* increases by *1 unit*.

As *precipitation* does not have a significant statistical value for the model (based on the regression coefficients formula), we considered it unfeasible to reliably plot residuals or conduct further analysis. Overall, there is no significant difference in comparison with Scenario 1.1. The model containing the *humidity* variable appears stronger than the one containing the *precipitation* variable.

4.2. Scenario 2: Dependence of car accidents on faulty infrastructure

For this scenario, we have used *car accidents* as a dependent variable, with *faulty lights* and *electricity charges* as independent variables.

4.2.1. Preliminary results

From Fig. 7, we can infer a general *positive correlation* between *faulty lights*, *electricity charges* and *car accidents* throughout the evaluated period, with peaks in activity starting early July and ending early-mid September.

We could give the same explanation for this trend as in the previous scenario. In general, days characterised by *peaks* of *faulty lights* and *electricity charges* also feature *peaks* in *car accidents*. The same occurs for the days exhibiting *dips* in both metrics. Overall, we can infer a general *positive correlation* between *faulty lights*, *electricity charges*, and *car accidents*.

Table 11
Regression statistics for Scenario 2.

Multiple R	0.68
R^2	0.462
Adjusted R^2	0.458
Standard error	1.41

Table 12
ANOVA table for Scenario 2.

	SS	MS	F	Sig. F
Regression	440.46	220.23	109.39	5.39E−35
Residual	511.36	2.01		
Total	951.82			

Table 13
Regression coefficients for Scenario 2.

	Coefficients	Standard error	t Stat	P-value
Intercept	0.3275576729	0.109528596	2.990613271	0.003057774143
Faulty lights	0.2690306152	0.09128012226	2.947307788	0.003503706361
Electricity charges	0.2979115757	0.02992479622	9.955341835	6.31E−20

4.2.2. Regression statistics and ANOVA

In Table 11, $R^2 = 0.462$ suggests a *direct* and *medium intensity* connection between the three variables. *Adjusted* $R^2 = 0.458$ shows that 45.8% of the variation of the *car accidents* variable can be explained by the variation of the *faulty lights* and the *electricity charges* variables. The estimated value of the *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}} = 0.6802$ indicates a strong link between the model variables. From Table 12, based on the small value of the f-test = 5.39E−35, we reject the H_0 hypothesis for both 1% and 5% risk. It can be stated that the model significantly explains the connection between the variables, and a linear statistical dependence exists between the variables.

4.2.3. Regression coefficients and variability

Table 13 dictates the multiple linear regression model equation: $Y = 0.32755 + 0.26903 * x_1 + 0.2979 * x_2$, where x_1 = *faulty lights*, and x_2 = *electricity charges*. All parameters are statistically significant since, after computing the *t-tests*, we reject H_0 due to $\sigma < \alpha$. For β_0 (*car accidents*), $\frac{b_0}{s_{\hat{\beta}0}} = 0.3277/0.1095 = 2.992$. For β_1 (*faulty*

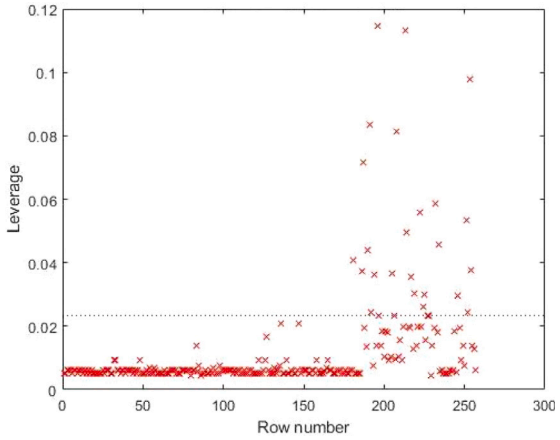


Fig. 8. Case order plot of leverage — Scenario 2.

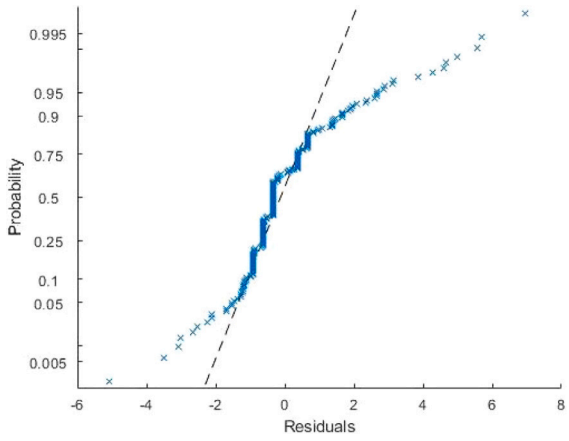


Fig. 9. Probability plot of residuals — Scenario 2.

lights), $\frac{b_1}{s_{\hat{\beta}1}} = 0.3277/0.1095 = 2.992$. For β_2 (*electricity charges*), $\frac{b_2}{s_{\hat{\beta}2}} = 0.3277/0.1095 = 2.992$. For β_0 (*car accidents*), the average is 0.3275 when the other variables are simultaneously equal to 0. For β_1, it increases by 0.269 when *faulty lights* increase by *1 unit*. For β_2, it increases by 0.297 when the *electricity charges* increase by *1 unit*.

4.2.4. Residual and outliers analysis

The analysis of residuals in Fig. 8 presents certain data portions with higher leverage than the rest, suggesting the presence of outliers, also confirmed by the standard distribution plot of the residuals in Fig. 9. A closer examination suggests that the residuals do not feature a normal distribution but rather a closer resemblance to a heavy-tail. Although not a sufficient condition to invalidate the model, it presents the possibility of uncertainty within it. Overall, the model exhibits a reasonable degree of accuracy despite noticeable outliers present in the data. The model suggests that variations in the independent variable *faulty lights* have a higher impact than *electricity charges* on the dependent variable *car accidents*.

4.3. Scenario 3: Dependence of thefts on gatherings and temperature

In this scenario, we used *thefts* as a dependent variable, with *gatherings* and *temperature* as independent variables.

4.3.1. Preliminary results

From Fig. 10, we can infer a general *positive correlation* between *gatherings* and *thefts* throughout the evaluated period of 2020. We can observe a peak in activity starting from early July and spanning through early-mid September, which could be explained by the summer period, generally accompanied by more gatherings than the rest of the year, as well as higher temperatures. In general, days with *peaks* in *gatherings* also feature *peaks* in *thefts*. The same is true for days exhibiting *dips* in both metrics within a period that also coincides with higher *temperatures*. Overall, we can infer a general *positive correlation* between *gatherings*, *thefts*, and *temperature*.

Table 14
Regression statistics for Scenario 3.

Multiple R	0.80
R^2	0.64
Adjusted R^2	0.64
Standard error	2.21

Table 15
ANOVA table for Scenario 3.

	SS	MS	F	Sig. F
Regression	2290.21	1145.1	233.8	2.50E−5
Residual	1243.58	4.8		
Total	3533.79			

Table 16
Regression coefficients for Scenario 3.

	Coefficients	Standard error	t Stat	P-value
Intercept	−1.038999876	0.3673023976	−2.828731539	0.005046172269
Gatherings	0.3868833403	0.02363598937	16.3684005	1.32E−41
Temperature	0.1479463585	0.03215363987	4.601232057	0.000006635872262

4.3.2. Regression statistics and ANOVA

In Table 14, $R^2 = 0.64$ shows a direct link between the four variables of strong intensity. However, *Adjusted* R^2 0.64 means that 64% of the number of thefts can be explained by the variations of the *gatherings* and *temperature* variables.

The value of the *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}} = 0.805$ indicates a strong link between the variables. The small value of the f-test = 2.4954E−58 in Table 15 leads us to reject the H_0 hypothesis for both 1% and 5% risk. We could state that the model significantly explains the connections between variables, exhibiting a linear statistical dependence.

4.3.3. Regression coefficients and variability

Table 16 dictates the multiple linear regression model equation: $Y = -1.038 + 0.3868 * x_1 + 0.147 * x_2$, where x_1 = *gatherings* and x_2 = *temperature*. All parameters are statistically significant since, after computing the *t-tests*, we reject H_0 due to $\sigma < \alpha$ in all $t - test$ values. For β_0 (*thefts*), $\frac{b_0}{s_{\hat{\beta}0}} = -1.038/0.36 = -2.82$ For β_1 (*gatherings*), $\frac{b_1}{s_{\hat{\beta}1}} = 0.386/0.023 = 16.38$. For β_2 (*temperature*), $\frac{b_2}{s_{\hat{\beta}2}} = 0.1479/0.032 = 4.60$. For β_0 (*thefts*), the average is equal to −1.03 when all other variables are simultaneously equal to 0. For β_1, it increases by 0.38 when *gatherings* increase by *1 unit*. For β_2, it increases by 0.14 when the *temperature* increases by *1 unit*.

4.3.4. Residual analysis

The distribution of the residuals in Fig. 11 exhibits a reasonable spread. The higher values suggest outliers present in the data, also observable in the standard probability plot (Fig. 12), where deviation from the normal distribution is present at the lower and upper bands. Overall, the model is accurate despite noticeable outliers in the data. The coefficients formula confirms that both independent variables have a positive effect on the dependent variable *thefts*, while *gatherings* has a more pronounced influence than *temperature*.

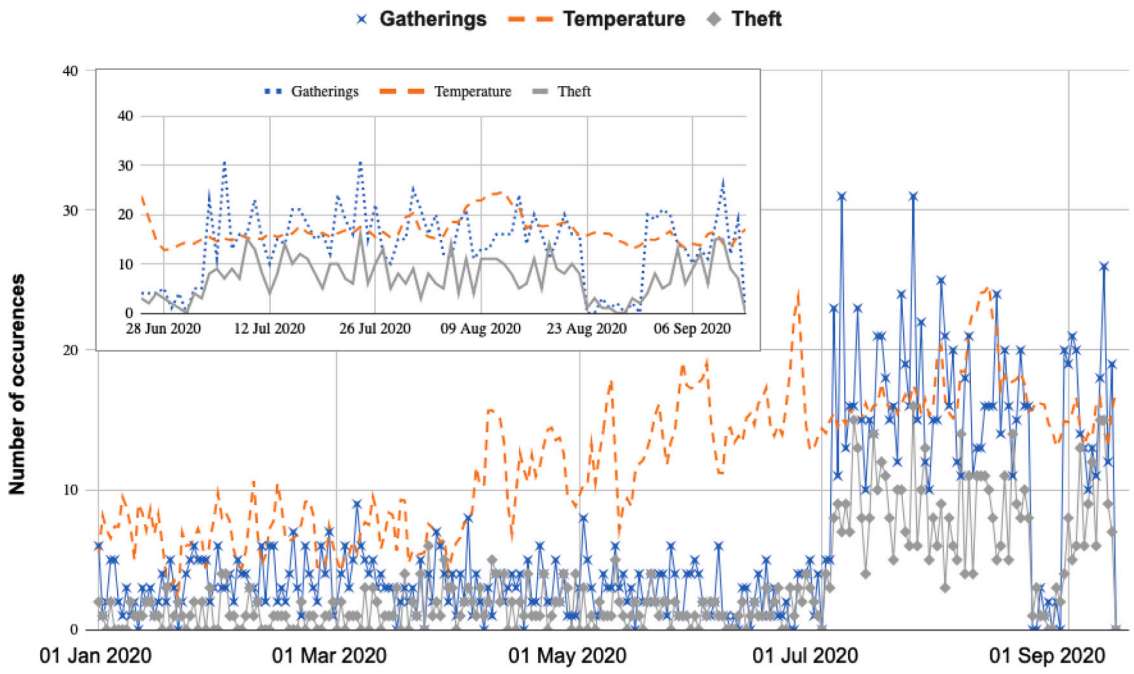


Fig. 10. Preliminary results of social media analysis for Scenario 3 — Smart City event occurrences.

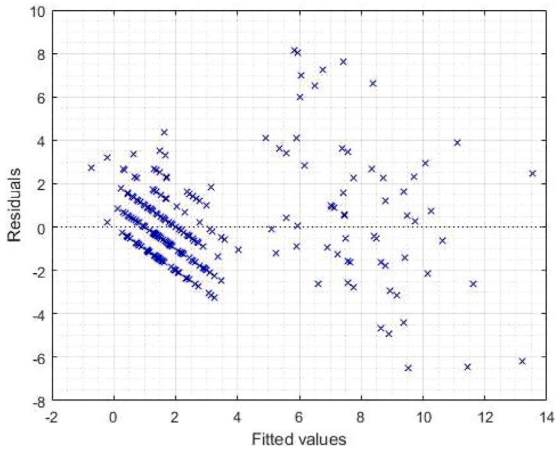


Fig. 11. Residuals vs. fitted values — Scenario 3.

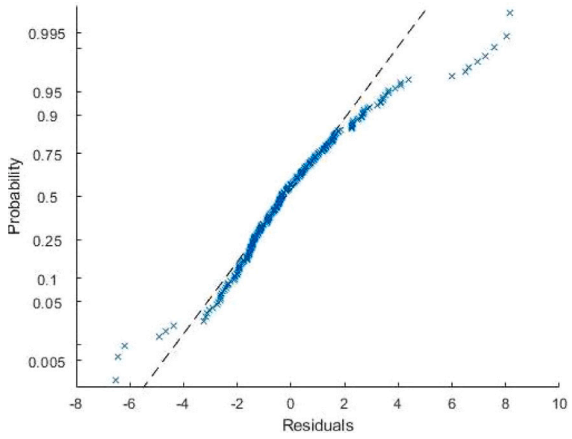


Fig. 12. Normal probability of residuals — Scenario 3.

4.4. Scenario 4: Dependence of citizen satisfaction on gatherings, queues and electricity charges

In the first part of the current scenario analysis (Scenario 4.1), we have used *positive sentiments* as a dependent variable and *gatherings*, *queues* and *electricity charges* as independent variables. In the second part (Scenario 4.2), we have replaced the dependent variable with *negative sentiments*.

4.4.1. Preliminary results

When observing the preliminary results for Scenario 4 as a whole, Fig. 13 indicates a general *positive correlation* between the emotional predilection of citizens and *gatherings*, with *peaks* of *positive sentiments* occurring in summer and in the proximity of periods characterised by low occurrences of *queues* and *electricity charges* respectively. Fig. 14 shows a *peak* of *positive sentiments* in the second half of summer (July 25th to August 30th), followed by a *dip* in the first part of autumn (September 1st–13th). This trend could be partially explained by the co-occurrence of COVID restrictions with the start of the autumn season, featuring higher levels of humidity and precipitation. We can also observe small *peaks* of *positive* emotional predilection in April and May 2020, periods with fewer *queues* and *electricity charges*. Marginal exceptions are the days characterised by *peaks* in queues and *dips* in *gatherings*, which are also characterised by *dips* in *positive sentiments* (the days around the 1st of May and the days in late June).

Overall, we can infer a *negative correlation* between the independent variables (*queues* and *electricity charges*) and the independent variable (*positive sentiments*).

4.4.2. Regression statistics and ANOVA — Scenario 4.1

In Table 17, the first indicator of a valid and significant model is $R^2 = 0.92$, which shows a robust and direct link between variables. A more detailed analysis may raise the issue of non-compliance with the hypothesis regarding the modelling error, and the co-linearity phenomenon might occur.

Adjusted $R^2 = 0.921$ means that 92.1% of the variation of the positive variable can be explained by the simultaneous variation of the independent variables. The estimated value of the *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}} = 0.956$ indicates a strong link between the model variables.

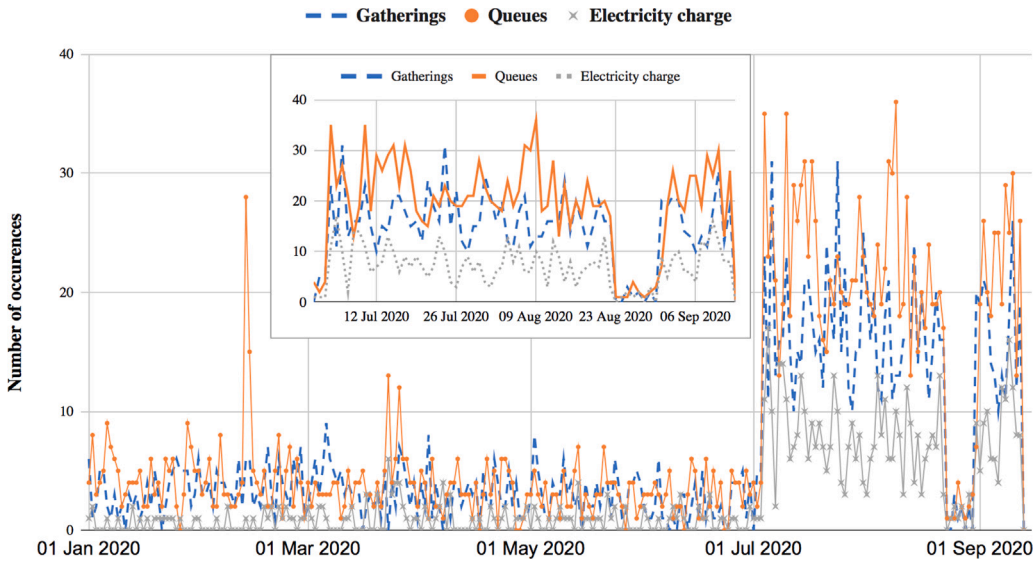


Fig. 13. Preliminary results of social media data analysis for Scenario 4 — Smart City event occurrences.

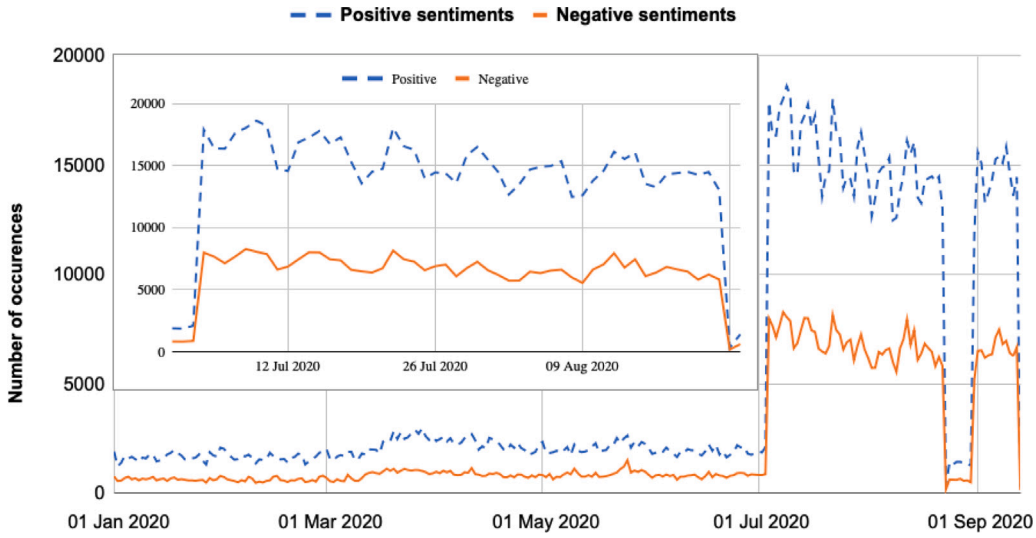


Fig. 14. Preliminary results of social media data analysis for Scenario 4 — Citizen satisfaction a Smart City.

Table 17
Regression statistics for Scenario 4.1.

Multiple R	0.96
R^2	0.92
Adjusted R^2	0.92
Standard error	1615.42

Table 18
ANOVA table for Scenario 4.1.

	SS	MS	F	Sig. F
Regression	7 825 425 722	2 608 475 241	999.57	6.31E−140
Residual	660 225 378.3	2 609 586.47		
Total	8 485 651 100			

In Table 18, the small f-test = 6.306E−140 leads us to reject the H_0 hypothesis for both a 1% and a 5% risk. It can be stated that the model significantly explains the connection between the variables, and a *linear statistical dependence* exists between the variables.

Table 19
Regression coefficients for Scenario 4.1.

	Coefficients	Standard error	t Stat	P-value
Intercept	−69.3800431	142.6485889	−0.4863703428	0.6271256406
Gatherings	339.1351667	29.05710087	11.67133529	1.82E−25
Queues	220.5609509	24.23102201	9.102420479	2.73E−17
Electricity charges	465.1377969	48.73416334	9.54438868	1.23E−18

4.4.3. Regression coefficients and variability — Scenario 4.1

Table 19 dictates the multiple linear regression model equation: $Y = -69.380 + 339.135 * x_1 + 220.5 * x_2 + 465.13 * x_3$, where x_1 = *gatherings*, x_2 = *queues*, and x_3 = *electricity charges*. For the parameter *positive sentiments*, we do not reject H_0, and conclude that the parameter β_0 is not statistically significant, since $\frac{b_0}{s_{\hat{\beta}0}} = -69.38/142.6 = -0.48$ and $\sigma > \alpha$. All other parameters are statistically significant since, after computing *t-tests*, we reject H_0 due to $\sigma < \alpha$.

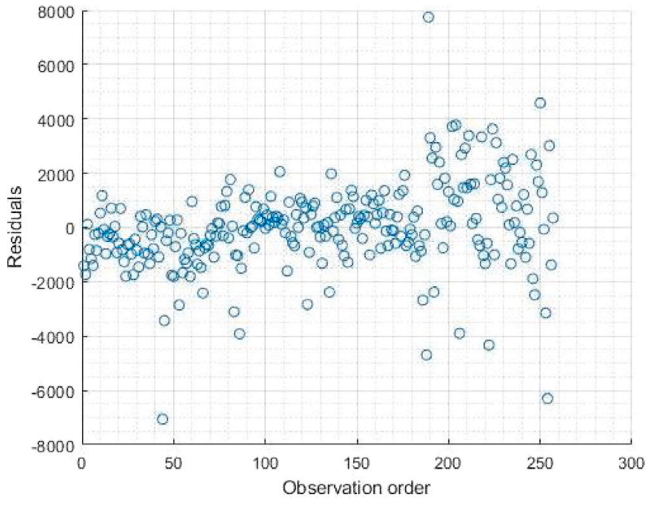


Fig. 15. Residuals vs. Observation Order Scenario 4.1.

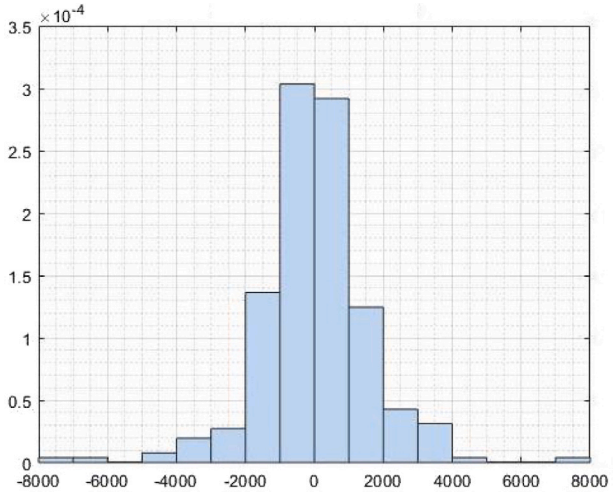


Fig. 16. Histogram of residuals for Scenario 4.1.

For β_1 (*gatherings*), $\frac{b_1}{s_{\hat{\beta}1}}$ = 339.13/29.05 = 11.67. For β_2 (*queues*), $\frac{b_2}{s_{\hat{\beta}2}}$ = 220.56/24.23 = 9.10. For β_3 (*electricity charges*), $\frac{b_3}{s_{\hat{\beta}3}}$ = 465.13/48.73 = 9.54. For β_0 (*positive sentiments*), the average is −69.38 when all other variables are simultaneously equal to 0. For β_1, it increases by 399.135 when the number of *gatherings* increases by *1 unit*. For β_2, it increases by 220.5 when *queues* increase by *1 unit*. For β_3, it increases by 465.137 when *electricity charges* increase by *1 unit*. The coefficients formula also confirms that *gatherings*, *queues* and *electricity charges* have a positive effect on the dependent variable *positive sentiments*, with *electricity charges* having a higher positive effect.

4.4.4. Residual analysis — Scenario 4.1

The plot in Fig. 15 illustrates the residual distribution against the observation order. The primarily linear upper and lower bands suggest a significant relationship between the independent and dependent variables. Outliers can be observed in the residuals' histogram (Fig. 16), which also shows that most residuals follow the normal distribution. Despite their noticeable presence observed around the range of 6000–8000, the model presents itself as accurate and representative.

Table 20
Regression statistics for Scenario 4.2.

Multiple R	0.95
R^2	0.91
Adjusted R^2	0.91
Standard error	774.07

Table 21
ANOVA table for Scenario 4.2.

	SS	MS	F	Sig. F
Regression	1 621 137 835	540 379 278.2	901.84	9.76E−13
Residual	151 596 626.4	599 196.15		
Total	1 772 734 461			

Table 22
Regression coefficients — Scenario 4.2.

	Coefficients	Standard error	t Stat	P-value
Intercept	−128.471671	68.35434741	−1.879495245	0.06132622005
Gatherings	148.6386076	13.92358089	10.67531469	3.31E−22
Queues	102.2776672	11.61102054	8.808671631	2.06E−16
Electricity charges	217.6853553	23.35243521	9.321741109	5.90E−18

4.4.5. Regression statistics and ANOVA — Scenario 4.2

This Section commences the second part of our scenario, where we considered *negative sentiments* as a dependent variable and *gatherings*, *queues* and *electricity charges* as independent variables.

The model with the *negative* dependent variable is almost as strong as the one with the *positive* dependent variables. From Table 20 we can observe a strong and positive correlation, where the value of *Adjusted* R^2, 91.3% of the variation of the *negative* variable can be explained by the variation of the independent variables in the model. The estimated value of the *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}}$ = 0.956 indicates a strong link between the model variables. The low value of f-test = 6.306E−140 in Table 21 leads us to reject H_0 for both 1% and 5% risk and state that the model significantly explains the connection between the variables, exhibiting a linear statistical dependence.

4.4.6. Regression coefficients and variability — Scenario 4.2

Table 22 dictates the multiple linear regression model equation: $Y = -128.47 + 148.63 * x_1 + 102.27 * x_2 + 217.68 * x_3$, where x_1 = *gatherings*, x_2 = *queues*, and x_3 = *electricity charges*.

For β_0 (*negative sentiments*), since $\frac{b_0}{s_{\hat{\beta}0}} = -128.47/68.35 = -1.87$ and $\sigma > \alpha$, we do not reject H_0, and conclude that the parameter β_0 is not statistically significant. All other parameters are statistically significant since, after computing *t-tests*, we reject H_0 due to $\sigma < \alpha$. For β_1 (*faulty lights*), $\frac{b_1}{s_{\hat{\beta}1}}$ = 148.63/13.92 = 10.6. For β_2 (*queues*), $\frac{b_2}{s_{\hat{\beta}2}}$ = 102.2/11.6 = 8.80. For β_3 (*electricity charges*), $\frac{b_3}{s_{\hat{\beta}3}}$ = 217.68/23.35 = 9.31.

For β_0 (*negative sentiments*), the average is −128.47 when all other variables are simultaneously equal to 0. For β_1, it increases by 148.63 when the number of gatherings increases by *1 unit*. For β_2, it increases by 102.27 when the number of queues increases by *1 unit*. For β_3, it increases by 217.68 when the *electricity charges* increase by *1 unit*. The coefficients formula also confirms that *gatherings*, *queues* and *electricity charges* have a positive effect on the dependent variable, with *electricity charges* having a higher positive effect.

4.4.7. Residual analysis — Scenario 4.2

Fig. 17 illustrates the residual distribution against the observation order, with mostly linear upper and lower bands suggesting a significant relationship between the independent dependent variables. Despite outliers present in the histogram of residuals (Fig. 18), a normal distribution (−3000–3000) indicates an accurate and representative model.

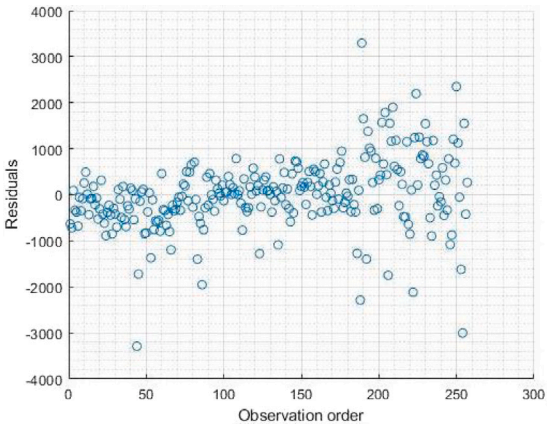


Fig. 17. Residuals vs. observation order — Scenario 4.2.

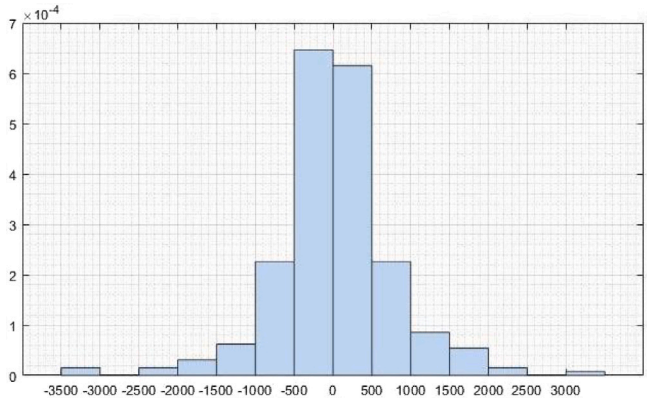


Fig. 18. Histogram of residuals — Scenario 4.2.

4.5. Scenario 5: Dependence of citizen satisfaction on environmental factors

In the first part of this scenario analysis, we considered *positive sentiments* as a dependent variable and *temperature*, *precipitation* and *humidity* as independent variables. In the second part, we replaced the dependent variable with *negative sentiments*.

4.5.1. Preliminary results

We can observe *peaks* of positive sentiments in the second half of summer (July 25th to August 30th), coupled with a *dip* in the first part of the autumn (September 1st–13th), which could be explained by the lock-downs co-occurring with the start of the autumn, with high *precipitation* and *humidity* (see Figs. 19 and 20).

As with other scenarios, we can observe small *peaks* of positive emotional predilection in April and May 2020, months with *higher temperature* and *lower precipitation*. Marginal exceptions are the days with *peaks* in precipitation and *dips* in temperature, which are also characterised by *dips* in positive sentiments (the days around the 1st of May and the days in late June). A general *positive correlation* emerges between the citizen satisfaction and weather, with *peaks* of positive sentiments occurring in summer and close to periods featuring low *precipitation* and *humidity*.

4.5.2. Regression statistics and ANOVA — Scenario 5.1

From Table 23, we can observe potential improvements for the model, since $R^2 = 0.364$ suggests a direct connection, but of low intensity: only 35.7% of the variation of *positive sentiments* can be explained by the variation of the three independent variables. *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}} = 0.603$ indicates an average link between

Table 23
Regression statistics for Scenario 5.1.

Multiple R	0.60
R^2	0.36
Adjusted R^2	0.35
Standard error	4616.48

Table 24
ANOVA table for Scenario 5.1.

	SS	MS	F	Sig. F
Regression	3 093 729 675	1 031 243 225	48.38	9.45E−25
Residual	5 391 921 425	21 311 942.39		
Total	8 485 651 100			

Table 25
Regression coefficients and variability — Scenario 5.1.

	Coefficients	Standard error	t Stat	P-value
Intercept	−12 262.06101	2713.834916	−4.518351847	0.000009571525336
Temperature	733.9261444	62.19320143	11.80074554	6.76E−26
Precipitation	−201.8790636	86.32112974	−2.338698117	0.02013027635
Humidity	120.0656664	32.63285245	3.679288123	0.0002857606397

the model variables. Observing Table 24, the low value of f-test = $9.4516E{-}25$ leads us to reject H_0, for both 1% and 5% risk. Therefore, the model significantly explains the connections between variables and exhibits a linear statistical dependence.

4.5.3. Regression coefficients — Scenario 5.1

Table 25 dictates the multiple linear regression model equation: $Y = -12262.061 + 733.92 * x_1 - 201.87 * x_2 + 120.06 * x_3$, where x_1 = *temperature*, x_2 = *precipitation*, and x_3 = *humidity*. All parameters are statistically significant since, after computing the *t-tests*, we reject H_0 due to $\sigma < \alpha$.

For β_0 (*positive sentiments*), $\frac{b_0}{s_{\hat{\beta}0}} = -12262/2713 = -4.518$. For β_1 (*temperature*), $\frac{b_1}{s_{\hat{\beta}1}} = 733/62 = 11.8$. For β_2 (*precipitation*), $\frac{b_2}{s_{\hat{\beta}2}} = -201/86 = -2.33$. For β_3 (*humidity*), $\frac{b_3}{s_{\hat{\beta}3}} = 120/32 = 3.67$.

For β_0 (*positive sentiments*) the average is −12262.061 when all other variables are simultaneously equal to 0. For β_1, it increases by 733.92 when *temperature* increases by *one degree*. For β_2, it decreases by −201.87 when *precipitation* increases by *1 unit*. For β_3, it increases by 120.06 when *humidity* increases by *1 unit*. The coefficients formula also confirms that both *temperature* and *humidity* have a positive effect on the dependent variable, *temperature* having the highest. Conversely, *precipitation* has a negative impact on *positive sentiments*.

4.5.4. Residual analysis — Scenario 5.1

The distribution of the residuals in Fig. 21 presents a clear pattern in the data. The downwards-pointing line distribution plot suggests either that a linear model is not the optimal representation of the data, or that clear dependencies are present in the model data, potentially explained by the physical interpretation of the variables: *humidity* could be a function of *temperature*, *precipitation*, or both.

The standard probability plot in Fig. 22 suggests that although noticeable outliers are present in the data, the overall distribution is normal. However, as the distribution of the residuals over the fitted values indicates, the independent variables are not, in fact, entirely independent.

4.5.5. Regression statistics and ANOVA — Scenario 5.2

This section commences the second part of our scenario analysis, where we considered *negative sentiments* as a dependent variable and *temperature*, *precipitation*, and *humidity* as independent variables.

From Table 26, we can infer that the model could benefit from improvements, as $R^2 = 0.37$ suggests a direct connection, but of low

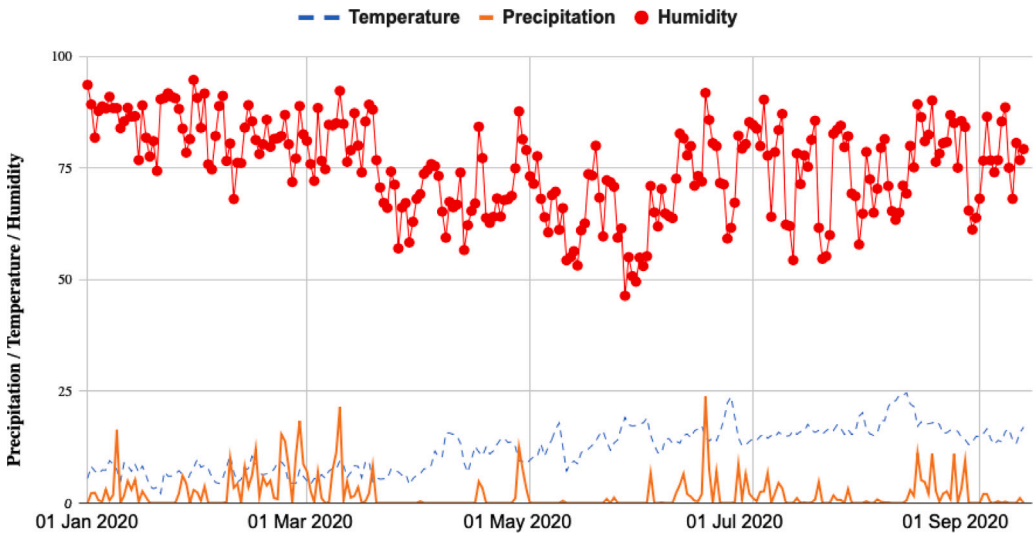


Fig. 19. Preliminary results of social media analysis for Scenario 5 — Smart City event occurrences.

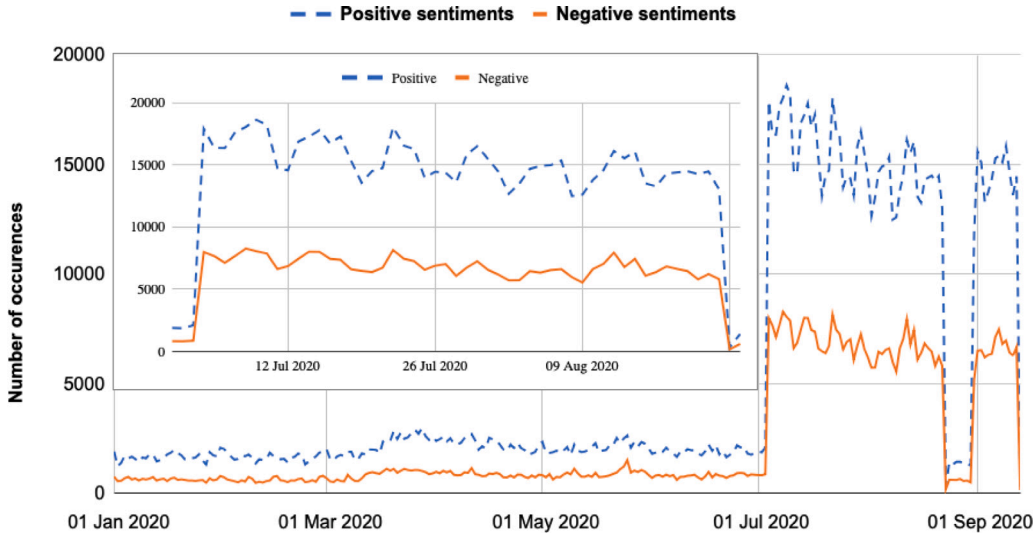


Fig. 20. Preliminary results of social media analysis for Scenario 5 — Citizen satisfaction in a Smart City context.

Table 26
Regression stat. — Scenario 5.2.

Multiple R	0.61
R^2	0.37
Adjusted R^2	0.37
Standard error	2086.85

Table 27
ANOVA table for Scenario 5.2.

	SS	MS	F	Sig. F
Regression	670 931 181.5	223 643 727.2	51.35	5.88E−26
Residual	1 101 803 280	4 354 953.674		
Total	1 772 734 461			

intensity, and only 37.1% of the variation of the dependent variable can be explained by the variation of the three independent variables. The value of the *Multiple Correlation Ratio* $\sqrt{\frac{ESS}{TSS}} = 0.615$ indicates an average intensity link between the model variables. The small value of the f-test = 5.879E−26 in Table 27 leads us to reject H_0 for both a 1% and 5% risk. It can be stated that the model significantly explains the connection between the variables, and a linear statistical dependence exists between the variables.

Table 28
Regression coefficients for Scenario 5.2.

	Coefficients	Standard error	t Stat	P-value
Intercept	−5871.495358	1226.771502	−4.786136087	0.000002893788971
Temperature	342.5242927	28.11403402	12.18339184	3.55E−27
Precipitation	−89.12472999	39.02090778	−2.284025028	0.02319907252
Humidity	55.7241524	14.75146966	3.777532252	0.0001975163616

4.5.6. Regression coefficients and variability — Scenario 5.2

Table 28 dictates the multiple linear regression model equation: $Y = -5871.49 + 342.52 * x_1 - 89.12 * x_2 + 55.72 * x_3$, where x_1 = *temperature*, x_2 = *precipitation*, and x_3 = *humidity*.

Table 28 also suggests that all parameters are statistically significant, since computing the *t-tests* leads us to reject H_0 due to $\sigma < \alpha$.

For β_0 (*negative sentiments*), $\frac{b_0}{s_{\hat{\beta}0}} = -5871.4/1226.7 = -4.78$. For β_1 (*temperature*), $\frac{b_1}{s_{\hat{\beta}1}} = 342.52/28.11 = 12.18$. For β_2 (*precipitation*), $\frac{b_2}{s_{\hat{\beta}2}} =$

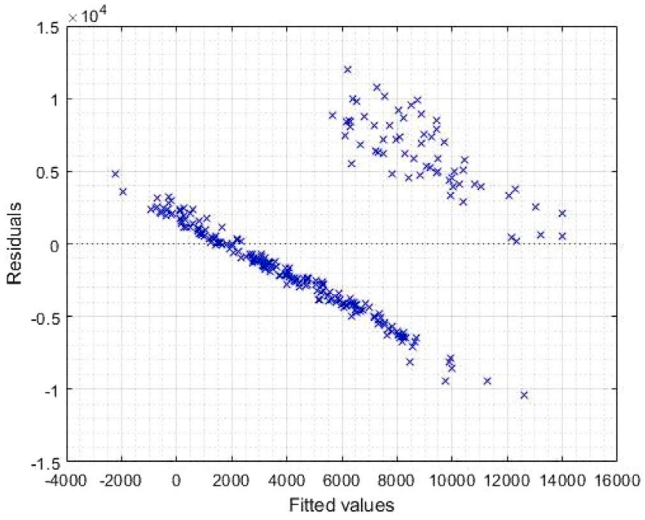


Fig. 21. Residual vs. fitted values — Scenario 5.1.

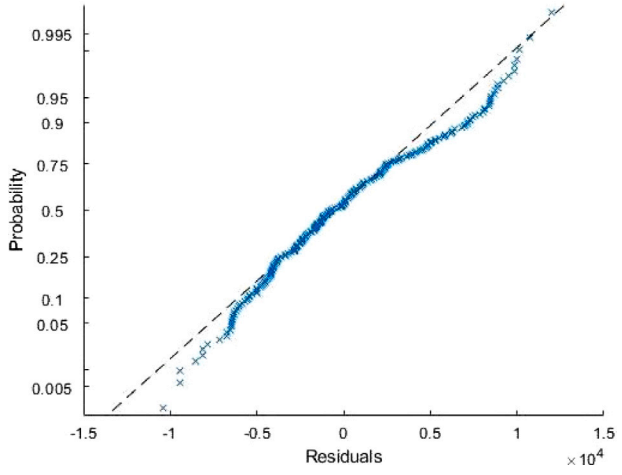


Fig. 22. Normal probability of residuals — Scenario 5.1.

$-89.12/39.02 = -2.2$. For β_3 (*humidity*), $\frac{b_3}{s_{\beta 3}} = 55.72/14.75 = 3.77$. For β_0 (*negative sentiments*), the average is −5871.4 when all other variables are simultaneously equal to 0. For β_1, it increases by 342.52 when temperature increases by one degree. For β_2, it decreases by −89.12 when precipitation increases by one unit. For β_3, it increases by 55.72 when humidity increases by *1 unit*.

The coefficients formula also confirms that *temperature* and *humidity* positively impact the dependent variable (*negative sentiments*), with *temperature* having a higher positive impact. We could also note that *precipitation* has a negative impact on the dependent variable (*negative sentiments*).

4.5.7. Residual analysis — Scenario 5.2

Fig. 23 represents the distribution plot of residuals over the fitted values. While appearing similar to Scenario 5.1, it also suggests the independent variables are not, in fact, entirely independent. Nevertheless, both models are promising. The independent variables largely explain the dependent variable's variation. We could notice a higher accuracy in the model featuring *positive sentiments* as a dependent variable than its counterpart. Part of future work plans is to treat the co-linearity problem while testing the hypothesis of errors.

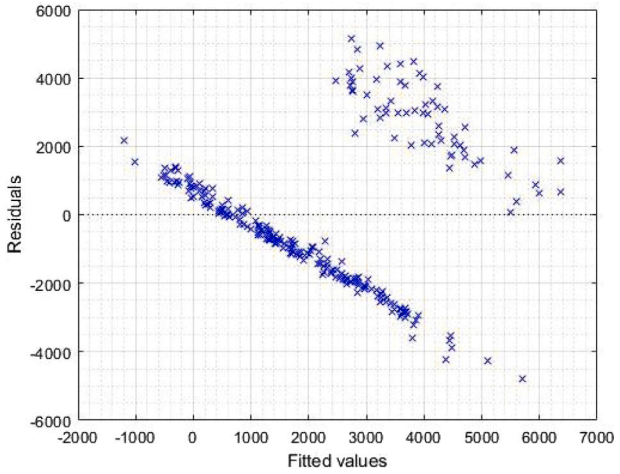


Fig. 23. Plot of residuals vs. fitted values — Scenario 5.2.

5. Discussion

This manuscript demonstrates that social media data streams represent a significant source of meaningful insight for adequate detection of events in a smart city context, helping achieve sustainability. Leveraging the power of our selected combination of Machine Learning and Natural Language Processing techniques (*AWD-LSTM* and *ULMFiT*) with an accuracy of 88.5% on our selected dataset (3% higher than the NLP techniques used by other researchers), we met our objectives. We addressed the positive research question by detecting real-time events broadcast by citizens acting as active social sensors, and matching the bespoke *Taxonomy*'s risk categories. Notwithstanding that it was derived from Coburn et al. (2014), the context-sensitive features are retained, and as such, its scalability. As elaborated in the NeOn methodology (Suárez-Figueroa, Gómez-Pérez, & Fernández-López, 2012), the taxonomy could be extended using competency questions to derive the key concepts forming the basis of an ontology.

We fulfilled our second research objective by converting raw data into a quantifiable format, facilitating manual and automatic detection of patterns and trends. In Section 4 we presented preliminary trends in *positive* and *negative* emotional predilection. Significant fluctuations were observed to be associated with variations in weather and occurrence rates of certain event types, such as *gatherings*, *thefts* and *electricity charges*. The data layout also enabled us to meet our third research objective: to validate and quantify the strength of the relationships between citizen satisfaction, environmental factors, and co-occurring events. While some scenarios exhibited more substantial regression statistics indicated by their Adjusted R^2 values above 90% (Scenario 4), all models provided valuable insights, as highlighted by their R^2 above 60% in all cases. As models were broken down into constituent variables, we have identified that all regression models contained at least one variable below the *0.05* significance threshold of the $f-test$.

The two scenarios concerning *car accidents* highlighted the potentially devastating effects that unfavourable environmental factors, coupled with instances of *congestion* and *faulty lights* might have. The *congestion* and *faulty lights* variables were assessed as highly significant in determining the number of *car accidents*, followed by *electricity charges* ranking second in correlation intensity. However, *humidity* was determined to have the least positive effects. Therefore, the message conveyed to authorities is that ensuring robust controls of *faulty lights*, *electricity charges*, and *queues* (related to both cars and pedestrians) could reduce *car accidents*, independently of weather conditions.

As for the scenario analysing the dependence of *thefts* on *gatherings* and *temperature*, the high Adjusted R^2 value of 64% indicated a direct link between the three variables. While highlighting the increased

crime rate at higher temperatures in the summer, the residual analysis concluded that *gatherings* have the most substantial positive effect on *thefts*. These results should encourage the authorities to deploy adequate resources during periods of peak population density in public settings, such as music festivals and cultural events during the summer.

The final two scenarios focused on analysing the citizen satisfaction levels at times of co-occurring events (such as *gatherings*, *queues* and *electricity charges*), while also considering environmental factors (*temperature*, *precipitation* and *humidity*).This scenario exhibited the strongest regression model out of all those in the *Taxonomy*, with an Adjusted R square of *92.1%*. We determined the *positive* emotional predilection to be most negatively correlated with *electricity charges*. Despite outliers identified by the histogram of residuals, the residual distribution against the observation order indicates the model is still statistically significant.

In relation to weather, the coefficients formula suggested that both *temperature* and *humidity* positively impact the dependent variable (*negative sentiments*). While *temperature* has the highest positive effect, *precipitation* has a *negative* impact on *positive sentiments*. The model employing the *positive sentiments* variable was less conclusive, suggesting a steep curve in *negative* emotional predilection due to weather fluctuations. The scenario emphasises the detrimental effects of the less favourable weather conditions (i.e., days characterised by *high precipitation*, *high humidity*, and *low temperatures*, coupled with reduced opportunities for *social gatherings*) on citizens' mental health. These findings could motivate the authorities to invest more resources into a sustainable strategy for socialising during times of adverse weather conditions, as well as to increase traffic fluidity, both on pavements and roads, in order to avoid *queues* adding to the *negative* emotional predilection.

The difference between the timestamps of an exact moment of an event occurring in the real world and the moment when our data harvesting and classifying engines detect the event is another known limitation of our research. The authors will address this limitation as part of their planned future work by linking the detected data with third-party data sources, such as IoT devices, sensors, and intelligence reports. Subject to building trust relationships with local authorities and law enforcement agencies, we could also link our data with third-party intelligence reports. In addition to enabling us to assign a "confidence index" to each of our detected events, in cases where third party intelligence emerges before an event is detected, we could factor this piece of data into a bespoke classification algorithm to improve accuracy and timely detection.

Another notable limitation relates to instances where the sentiment classifier is not entirely correct due to either logical fallacies in processing the natural language or subjective interpretation. In the present version of our implementation, in the absence of a manual review, these instances are difficult to isolate or statistically eliminate as outliers. In the future, we plan to implement additional features to the detection engine, which would also consider other indicators for emotional predilection, such as non-verbal language, sarcasm, and quotes from movies, books or other third-party sources. As a result of this inaccuracy, some of the risk assessments conducted at the post-occurrence stage, which take citizen satisfaction into account, might be rendered inaccurate and require further research into the optimum mitigation approach.

Finally, certain parts of our research raise privacy-related ethical concerns due to the potentially personally identifiable data collected. Devising a schema matching event categories to anonymised metrics (e.g., standard consumer profiles) would divert from the original intent of using social media as a primary data source. However, increased transparency into data usage, complementing the existing statutory wordings in privacy policies, would strengthen citizen trust and improve compliance with data processing regulations such as GDPR (General Data Protection Regulations).

6. Conclusions

This manuscript presents a framework for detecting events and analysis for smart cities with a subsequent citizen satisfaction overview of the emotional predilection of residents and transients defined by upstream research and initiates an appropriate response. The collected sentiment analysis and event detection data provide a snapshot of the events in a city at a given time. As such, it can be a valuable tool informing decision-making. Although the captured data sample analysed within this manuscript is partly stationary, the same research methodology could be replicated on real-time data or pre-selected time intervals of particular significance. The selected sample population is also statistically significant, as confirmed by the applied Multiple Regression Analysis statistical models. This research enabled us to comprehend the structure of a contemporary city and apply that knowledge to developing our *Taxonomy* of scenarios and events. In fulfilling our research objectives, we leveraged automated systems to aggregate and process data from social media with outputs reporting preliminary trends and Multiple Regression Analysis statistics.

Unlike a more general, all-encompassing approach, our results present a unique set of relationships created by the very functional core of any urban settlement: the citizens. Longitudinal satisfaction analysis and event classification enabled us to compare independent variables over time, while MRA produced quantitative results that validated the previously identified relationships. Therefore, the findings of our research, and any new findings produced by our proof-of-concept application, could be confidently utilised by authorities to make informed decisions about the development of smarter cities based on historical events and their own citizens as active agents within the urban environment, as opposed to being passive recipients of top-down management regimes. As a result, the city can meet the needs of its citizens through a bottom-up approach, which is highly relevant in the current complex and uncertain context posed by the pandemic and the climate agenda.

The rigour of our research is demonstrated by the multi-phase methodology utilised to firstly identify and then validate meaningful relationships between events, sentiments and weather. Our smart city scenarios, selected based on their relevance empirically established European Commission projects, were found to be statistically significant, featuring direct links of medium and strong intensity between variables, featuring at least one variable below the *0.05* significance threshold. The outliers in the residual plots could be explained through comparisons with preliminary results for each event category, pointing to fluctuations in COVID restrictions and environmental factors.

All in all, our research confirms that social media represents a reliable source of information informing smart city decision-making. However, in a real-world scenario, we acknowledge that a tight, multidisciplinary collaboration of specialists such as engineers, computer scientists and industry leaders would be necessary to achieve the most compelling results tailored to the specifics of each smart city ecosystem. We consider our research to make a significant step forward in the rapidly evolving field of applying information science techniques in the context of smart city communities, where the potential of leveraging citizens as social sensors and data broadcasters appears limitless.

CRediT authorship contribution statement

Andrei Hodorog: Designed the experiments, Methodology framework, Analysis tools, Analysed the data, Performed the experiments, Wrote the manuscript. **Ioan Petri:** Designed the experiments, Methodology framework, Analysis tools, Advised on the smart city applicability scenarios, Wrote the manuscript. **Yacine Rezgui:** Designed the experiments, Methodology framework, Analysis tools, Advised on the smart city applicability scenarios, Wrote the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to acknowledge the financial support from the EPSRC SemanticLCA Project (ref. EP/T019514/1). The authors would also like to acknowledge Augustin Arnault for his valuable insights and contributions to the manuscript.

References

Baculakova, K., et al. (2020). Selected aspects of smart city concepts: Position of bratislava. *Theoretical and Empirical Researches in Urban Management*, *15*(3), 68–80. http://dx.doi.org/10.1080/13604810802479126.

Bellini, E., Bellini, P., Cenni, D., Nesi, P., Pantaleo, G., Paoli, I., et al. (2021). An IoE and big multimedia data approach for urban transport system resilience management in smart cities. *Sensors*, *21*(2), 435. http://dx.doi.org/10.3390/s21020435.

Briskilal, J., & Subalalitha, C. (2022). Classification of idiomatic sentences using AWD-LSTM. In *Expert clouds and applications* (pp. 113–124). Springer, http://dx.doi.org/10.1007/978-981-16-2126-0_11.

Cardiff City Council (2020). Cardiff smart city roadmap. URL: https://www.smartcardiff.co.uk/roadmap/.

Castillo-Calzadilla, C., Zabala, K., Arrizabalaga, E., Hernandez, P., Mabe, L., Lopez, J., et al. (2021). The opportunity for smart city projects at municipal scale: Implementing a positive energy district in zorrozaurre. *Ekonomiaz*.

Cito, J., Schermann, G., Wittern, J. E., Leitner, P., Zumberi, S., & Gall, H. C. (2017). An empirical analysis of the docker container ecosystem on github. In *2017 IEEE/ACM 14th international conference on mining software repositories (MSR)* (pp. 323–333). IEEE.

Coburn, A. W., Bowman, G., Ruffle, S. J., Foulser-Piggott, R., Ralph, D., & Tuveson, M. (2014). A taxonomy of threats for complex risk management. *Cambridge Risk Framework Series*.

Colding, J., Colding, M., & Barthel, S. (2020). The smart city model: A new panacea for urban sustainability or unmanageable complexity? *Environment and Planning B: Urban Analytics and City Science*, *47*(1), 179–187. http://dx.doi.org/10.1177/2399808318763164.

Croci, E., & Molteni, T. (2021). Business models for smart city solutions: An overview of main archetypes. *International Journal of Urban Planning and Smart Cities (IJUPSC)*, *2*(2), 94–109. http://dx.doi.org/10.4018/IJUPSC.2021070106.

Dimeski, B., Memeti, M., & Bogdanoska-Jovanoska, M. (2019). International cooperation of the city of skopje: Projects for accelerating smart city developments. *Smart Cities and Regional Development (SCRD) Journal*, *3*(1), 89–101.

Du, J., Zhu, Q., Shi, Y., Wang, Q., Lin, Y., & Zhao, D. (2020). Cognition digital twins for personalized information systems of smart cities: Proof of concept. *Journal of Management in Engineering*, *36*(2), Article 04019052. http://dx.doi.org/10.1061/me.2020.36.issue-2;ctype:string:Journal.

Faelens, L., Hoorelbeke, K., Soenens, B., Van Gaeveren, K., De Marez, L., De Raedt, R., et al. (2021). Social media use and well-being: A prospective experience-sampling study. *Computers in Human Behavior*, *114*, Article 106510. http://dx.doi.org/10.1016/j.chb.2020.106510.

Fasnacht, L. (2018). Mmappickle: Python 3 module to store memory-mapped numpy array in pickle format. *Journal of Open Source Software*, *3*(26), 651.

Gabaldón Moreno, A., Vélez, F., Alpagut, B., Hernández, P., & Sanz Montalvillo, C. (2021). How to achieve positive energy districts for sustainable cities: A proposed calculation methodology. *Sustainability*, *13*(2), 710. http://dx.doi.org/10.3390/su13020710.

Gao, Y., Li, Y., Sun, Y., Cai, Z., Ma, L., Pustišek, M., et al. (2022). IEEE access special section: Privacy preservation for large-scale user data in social networks. *IEEE Access*, *10*, 4374–4379. http://dx.doi.org/10.1109/ACCESS.2020.3036101.

Girardi, P., & Temporelli, A. (2017). Smartainability: a methodology for assessing the sustainability of the smart city. *Energy Procedia*, *111*, 810–816.

Hodorog, A., Petri, I., Rezgui, Y., & Hippolyte, J. L. (2021). Building information modelling knowledge harvesting for energy efficiency in the construction industry. *Clean Technologies and Environmental Policy*, *23*(4), 1215–1231. http://dx.doi.org/10.1007/s10098-020-02000-z.

Howard, J., & Gugger, S. (2020). Fastai: a layered API for deep learning. *Information*, *11*(2), 108.

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146.

Ionescu, V. M. (2015). The analysis of the performance of rabbitmq and activemq. In *2015 14th roedunet international conference-networking in education and research (RoEduNet NER)* (pp. 132–137). IEEE.

Kalinin, M., Krundyshev, V., & Zegzhda, P. (2021). Cybersecurity risk assessment in smart city infrastructures. *Machines*, *9*(4), 78. http://dx.doi.org/10.3390/machines9040078.

Kim, D., & Kim, S. (2022). Role and challenge of technology toward a smart sustainable city: topic modeling, classification, and time series analysis using information and communication technology patent data. *Sustainable Cities and Society*, Article 103888.

Kramer, O. (2016). Scikit-learn. In *Machine learning for evolution strategies* (pp. 45–53). Springer.

Kummitha, R. K. R., & Crutzen, N. (2019). Smart cities and the citizen-driven internet of things: A qualitative inquiry into an emerging smart city. *Technological Forecasting and Social Change*, *140*, 44–53. http://dx.doi.org/10.1016/j.techfore.2018.12.001.

Lamba, M., & Madhusudhan, M. (2018). Application of sentiment analysis in libraries to provide temporal information service: a case study on various facets of productivity. *Social Network Analysis and Mining*, *8*(1), 63. http://dx.doi.org/10.1007/s13278-018-0541-y.

Li, Y., Rezgui, Y., & Kubicki, S. (2020). An intelligent semantic system for real-time demand response management of a thermal grid. *Sustainable Cities and Society*, *52*, Article 101857.

Lytras, M. D., Visvizi, A., & Jussila, J. (2020). Social media mining for smart cities and smart villages research. http://dx.doi.org/10.1007/s00500-020-05084-3.

Malche, T., Maheshwary, P., & Kumar, R. (2019). Environmental monitoring system for smart city based on secure internet of things (IoT) architecture. *Wireless Personal Communications*, *107*(4), 2143–2172. http://dx.doi.org/10.1007/s11277-019-06376-0.

Merity, S., Keskar, N. S., & Socher, R. (2017). Regularizing and optimizing LSTM language models. arXiv preprint arXiv:1708.02182.

Morishita-Steffen, N., Alberola, R., Mougeot, B., Vignali, E., Wikström, C., Montag, U., et al. (2021). Smarter together: Progressing smart data platforms in Lyon, Munich, and Vienna. *Energies*, *14*(4), 1075.

Olivadese, R., Alpagut, B., Revilla, B. P., Brouwer, J., Georgiadou, V., Woestenburg, A., et al. (2021). Towards energy citizenship for a just and inclusive transition: Lessons learned on collaborative approach of positive energy districts from the eu horizon2020 smart cities and communities projects. In *Multidisciplinary Digital Publishing Institute Proceedings, Vol. 65–1* (p. 20).

Rahimi-Golkhandan, A., Aslani, B., & Mohebbi, S. (2021). Predictive resilience of interdependent water and transportation infrastructures: A sociotechnical approach. *Socio-Economic Planning Sciences*, Article 101166. http://dx.doi.org/10.1016/j.seps.2021.101166.

Rehman, A. U., Naqvi, R. A., Rehman, A., Paul, A., Sadiq, M. T., & Hussain, D. (2020). A trustworthy siot aware mechanism as an enabler for citizen services in smart cities. *Electronics*, *9*(6), 918. http://dx.doi.org/10.3390/electronics9060918.

Russo, V. (2019). Digital economy and society index (DESI). European guidelines and empirical applications. *Qualitative and Quantitative Models in Socio-Economic Systems and Social Work*, *208*, 427. http://dx.doi.org/10.1007/978-3-030-18593-0_31.

Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S.-g., Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1–34. http://dx.doi.org/10.1186/s13673-019-0205-6.

Sharida, A., Hamdan, A., & Mukhtar, A.-H. (2020). Smart cities: The next urban evolution in delivering a better quality of life. In *Toward social internet of things (SIoT): Enabling technologies, architectures and applications* (pp. 287–298). Springer, http://dx.doi.org/10.1007/978-3-030-24513-9_16.

Suárez-Figueroa, M. C., Gómez-Pérez, A., & Fernández-López, M. (2012). The NeOn methodology for ontology engineering. In *Ontology engineering in a networked world* (pp. 9–34). Springer.

Temeljotov Salaj, A., & Loewen, B. (2020). The next generation of smart citizens: Experiences and inspiration from the+ CityxChange project. In *7th international academic conference on places and technologies*. University of Belgrade, Faculty of Architecture Belgrade.

Wirtz, B. W., Göttel, V., Langer, P. F., & Thomas, M.-J. (2020). Antecedents and consequences of public administration's social media website attractiveness. *International Review of Administrative Sciences*, *86*(1), 38–61. http://dx.doi.org/10.1177/0020852318762310.

Xavier, C. C., & Souza, M. (2020). A basic approach for extracting and analyzing data from Twitter. In *Special topics in multimedia, IoT and web technologies* (pp. 185–211). Springer.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information Sciences*, *307*, 39–52.

Yang, J., Xiu, P., Sun, L., Ying, L., & Muthu, B. (2022). Social media data analytics for business decision making system to competitive analysis. *Information Processing & Management*, *59*(1), Article 102751. http://dx.doi.org/10.1016/j.ipm.2021.102751.

Yuan, Y., Lu, Y., Chow, T. E., Ye, C., Alyaqout, A., & Liu, Y. (2020). The missing parts from social media–enabled smart cities: Who, where, when, and what? *Annals of the American Association of Geographers*, *110*(2), 462–475. http://dx.doi.org/10.1080/24694452.2019.1631144.

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., et al. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. *Information Processing & Management*, *56*(6), Article 102097. http://dx.doi.org/10.1016/j.ipm.2019.102097.