Periñán-Pascual, Carlos (2023) From Smart City to Smart Society: A quality-of-life ontological model for problem detection from user-generated content. Applied Ontology 18 (3), 263-306.
DOI 10.3233/AO-230281
IOS Press

From Smart City to Smart Society: A quality-of-life ontological model for problem detection from user-generated content

Carlos Periñán-Pascual*
Applied Linguistics Department, Universitat Politècnica de València, Paranimf 1, 46730 Gandia (Valencia), Spain
Email: jopepas3@upv.es

Abstract. Social-media platforms have become a global phenomenon of communication, where users publish content in text, images, video, audio or a combination of them to convey opinions, report facts that are happening or show current situations of interest. Smart-city applications can benefit from social media and digital participatory platforms when citizens become active social sensors of the problems that occur in their communities. Indeed, systems that analyse and interpret user-generated content can extract actionable information from the digital world to improve citizens' quality of life. This article aims to model the knowledge required for automatic problem detection to reproduce citizens' awareness of problems from the analysis of text-based user-generated content items. Therefore, this research focuses on two primary goals. On the one hand, we present the underpinnings of the ontological model that categorises the types of problems affecting citizens' quality of life in society. In this regard, this study contributes significantly to developing an ontology based on the social-sensing paradigm to support the advance of smart societies. On the other hand, we describe the architecture of the text-processing module that relies on such an ontology to perform problem detection, which involves the tasks of topic categorisation and keyword recognition.

Keywords**.** User-generated content, problem detection, text classification, keyword recognition, ontology

1. Introduction

Applications in smart cities rely on a large amount of real-time data that are typically collected from the large-scale deployment of heterogeneous physical sensors, such as those monitoring traffic congestions and road conditions, the consumption of utilities (e.g. power, water, and gas), ambient light and noise, etc. Smart cities contribute to digitising society by introducing innovative technologies. However, transforming smart cities into smart societies implies that technology solutions should be human-centred. For this reason, the concept of Smart City is moving towards the notion of Smart Society (Valkenburg et al., 2016), where citizens should be engaged to actively take part in creating a higher quality of life (QoL) not only for themselves but also for others. To this end, citizens should be provided with a space to participate and be involved:

> [...] if an open, multipurpose democratised platform is applied in the public domain, data can empower people to become active producers of societal value. (Valkenburg et al., 2016, p. 91)

* Corresponding author. E-mail: jopepas3@upv.es

Therefore, social-media platforms, which act as an interface for sharing thoughts and opinions, can make citizens become active participants in such smart societies. Regarding social-media platforms, a line of research that has gained popularity in the last decade is social sensing (Imran et al., 2015; Goswami & Kumar, 2016; Xu et al., 2018). An & Weber (2015) explained that various communities use the term *social sensing* differently. On the one hand, social sensing through physical devices involves physical sensors in mobile or wearable technology to infer social relationships and human activities (Madan et al., 2010; Gu et al., 2017). On the other hand, social sensing through social media involves the analysis of digital communications to detect real-world events or situations, where social-media users act as *sensors* (Sakaki et al., 2013; Musto et al., 2015; Arthur et al., 2018). In this article, social sensing refers to the latter concept, which includes collecting, analysing and interpreting user-generated content (UGC).

The notion of UGC is usually taken for granted. However, as it is an essential component in our research, it should be adequately defined. As no commonly agreed definition of UGC exists, Barbosa dos Santos (2022) performed a critical review of UGC to determine the scope of this concept. As a result, it can be concluded that UGC refers to the material (e.g. text, images, videos, and audio) created and published by users of Internet-based applications (e.g. web- and mobile-based platforms). This definition presents two aspects of UGC, i.e. creative effort and publication requirement. However, different approaches give a different value to each characteristic. In our research, the central characteristic of UGC is the possibility for users to publish content to others. When UGC is published, we say that it is contributed. In this regard, the contributor, who does not necessarily have to be the content creator, is the user who published the item. As described by Wyrwoll (2014), UGC typically refers to the content published on social media by users of platforms such as blogs, forums, location sharing and annotation platforms (e.g. Foursquare), media sharing platforms (e.g. Youtube), microblogs (e.g. Twitter), question and answer platforms (e.g. Blurtit), rating and review platforms (e.g. Tripadvisor), and social networks (e.g. Facebook). However, although social-media platforms are the primary source of UGC, this can also be published on digital participatory platforms (Falco & Kleinhans, 2018, p. 54):

> Digital Participatory Platforms (DPPs) are defined here as a specific type of civic technology explicitly built for participatory, engagement and collaboration purposes that allow for user generated content and include a range of functionalities (e.g. analytics, map-based and geo-located input, importing and exporting of data, ranking of ideas) which transcend and considerably differ from social media such as Social Networking Sites and Microblogging (Facebook, Twitter and Instagram).

Therefore, our approach to social sensing focuses on UGC, where (a) users play the role of citizens, and (b) the content describes events or situations perceived as problems that disrupt their QoL.

It is also worth mentioning that social sensors can supplement the measurements generated in smart cities by physical sensors. For example, data from physical sensors can identify what happens, but data from social sensors can shed light on why and how events emerge:

> physical sensors may identify lightly used bus routes by tracking passenger volume but these sensors cannot identify the underlying factors leading to underutilisation, which may include the route's cleanliness, its safety, and whether it reaches undesirable regions. (Doran et al., 2013, p. 1323)

Wang et al. (2019) explained that a "macroscope" can be developed as a new kind of instrument to view the physical and social reality as interpreted by the collective intelligence of social-media users,

thus providing valuable insights into our understanding of contemporary societies. To successfully address research challenges, the social-sensing field encompasses a synergy of several disciplines, such as social science, linguistics, knowledge engineering, information theory, and machine learning, to name a few.

In this context, this article results from the research conducted in the ALLEGRO project (Adaptive muLti-domain sociaL-media sEnsinG fRamewOrk), a system for the development of real-time multi-modal applications that can reconstruct the state of society as interpreted by the collective intelligence of social-media users. ALLEGRO consists of three core modules from which data fusion can be performed (i.e. text, audio, and image): DIAPASON (a unifieD hybrId ApProach to microtext Analysis in Social-media crOwdseNsing), SOUND (Social-media sOUnd aNalysis moDule), and ADAGIO (sociAl meDia imAGe analysIs mOdule). The goal of this article is twofold. First, we present the ontological model that categorises dimensions, domains, and types of problems affecting citizens' QoL in the Smart City framework. Second, we describe the role of this ontological model in DIAPASON, a workbench of exploration and experimentation with UGC written in English or Spanish to automatically detect a variety of problem types by integrating natural language processing (NLP), machine and deep learning, and knowledge engineering techniques. In this regard, the ontology is aimed at supporting DIAPASON in topic categorisation and keyword recognition, tasks that guided the development of this resource based on Ontology Design Patterns (ODP) (Hitzler et al., 2016; Krisnadhi and Hitzler, 2016). On the one hand, the structural ODP provides a taxonomic hierarchy organised into four levels of thematic granularity (i.e. problem realm, problem dimension, problem domain, and problem type) over which topics can be discovered in text classification. On the other hand, the content ODP conceptualises the semantics of specific problem types, whose problem schemas (i.e. conceptual representations for problem-type understanding) help determine the descriptors in keyword recognition. Therefore, in the context of Smart Society, interpreting a given text message posted on a social media or digital participatory platform involves tagging the UGC with the topic of the problem and the keywords that summarise the message. In short, this research aims to model the knowledge required for QoL diagnosis in general and in particular for automatic problem detection to reproduce citizens' awareness of problems in the Smart Society framework from the analysis of UGC.

The remainder of this article is organised as follows. Section 2 describes the most relevant works for this study around the central theme of Smart City. Section 3 presents the proposed research method, focusing on the underpinnings of the ontological model and the semantics of problem types. Section 4 examines how the DIAPASON text-processing module performs problem detection, which involves the tasks of topic categorisation and keyword recognition. Finally, our conclusions are drawn in Section 5.

2. Related work

2.1. Smart City: Definition and frameworks

In the era of the smart connected world, the concept *Smart City* usually refers to 'a city that manages, in an intelligent way, all its associated resources with the aim to enhance the quality of the services provided to citizens and to improve their quality of life' (Espinoza-Arias et al., 2019). Komninos et al. (2020, p. 12) denoted the city in the Internet era as a *cyber-physical city*, which 'has all features and qualities of the city plus the information, knowledge, and innovation generated by Information and Communication Technologies, IoT, and smart systems'. Indeed, they distinguished six types of cyber-physical cities, reserving the term Smart City to refer to cities in which e-services and data optimise city processes and infrastructures. As Smart City is a multidisciplinary concept, Ramaprasad et al. (2017) found more than thirty-six definitions. However, to provide a unified definition, they proposed a

framework based on six dimensions distributed into two aspects: *Smart*, where they considered the Structure, Function, Focus and Semiotics, and *City*, which is characterised by its Stakeholders and the Outcomes. Considering that each dimension consists of several components that can be encapsulated in the definition of Smart City, cities may be smart in different ways and degrees. According to this framework, for example, ALLEGRO is 'a system [Structure] to sense [Function] data [Semiotics] generated by citizens [Stakeholders] on social media and digital participatory platforms [Focus] for QoL diagnosis [Outcome]'.

Moreover, several frameworks have been proposed to characterise smart cities. One of the most prominent frameworks was presented by Giffinger et al. (2007), who defined smart cities along six key dimensions: Smart Economy (i.e. competitiveness), Smart Environment (i.e. natural resources), Smart Governance (i.e. public and social services and citizen participation), Smart Living (i.e. QoL), Smart Mobility (i.e. transport and communication infrastructure), and Smart People (i.e. social and human capital). In turn, these dimensions were broken down into thirty-one relevant factors that reflect the most critical aspects of every smart characteristic. Their framework was initially devised to rank medium-sized cities, where cities that perform well in these six dimensions are regarded as smart cities.

Govada et al. (2017) proposed enhancing Giffinger et al.'s (2007) Smart City framework by focusing on people, places and the planet in developing sustainable cities, thus resulting in three core values: Smart People, which comprises Smart Economy and Smart Governance, Smart Place, which includes Smart Mobility and Smart Living, and Smart Planet, which involves Smart Infrastructure and Smart Environment. Two significant changes were involved in this enhanced framework. On the one hand, Smart People is placed at a higher level to emphasise its importance as a primary element in developing truly smart cities. On the other hand, Smart Infrastructure is introduced, including physical (e.g. public works) and non-physical (e.g. the Internet) infrastructure.

Appio et al. (2019) distributed Giffinger et al.'s (2007) smart-city dimensions around Hutchison et al.'s (2011) five-level pyramid framework called "Intelligent Community Open Architecture" (i-COA). As shown in Fig. 1, the lower levels of this pyramid are the "hard" components of smart cities (i.e. physical), and the top levels represent the human-centric "soft" elements (i.e. social).

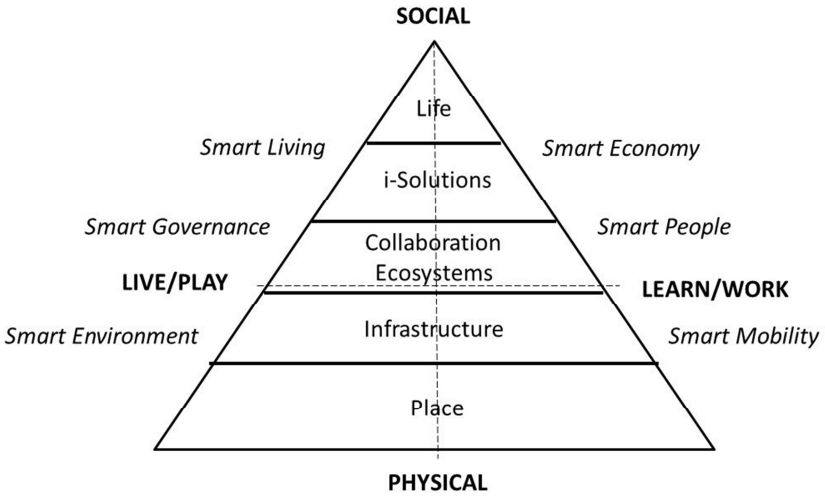

Fig. 1. Appio et al.'s (2019) proposal of Smart City framework.

2.2. Knowledge organisation systems for smart cities

Devices within smart cities are connected to share information and collaborate with other devices and with people so that the urban environment can be monitored. In this context, one of the major challenges is to integrate the vast quantity of data from heterogeneous sources and make sense of them so that actionable insights can be generated from data. Different varieties of data coexist in the smart connected world, which can be broadly classified into data from the physical world (i.e. sensor data) and data from the digital world (e.g. social data, web data, etc.). Moreover, to make sense of such data, we need to move from data to information to knowledge to wisdom (Ackoff, 1989). In other words, raw data, which result from observation, are useful only when transformed into information, which takes the form of descriptions about who, what, when, and how many; in turn, information should be transformed into knowledge, which can be inferred as know-how, i.e. what Jashapara (2005) described as actionable information. Therefore, organising knowledge is a critical issue in smart cities. In this regard, a knowledge organisation system (KOS) provides a framework for storing and organising data, information and knowledge about the world and thoughts for understanding, reasoning, discovery, retrieval, and many other purposes (Soergel, 2008). For example, a KOS can be used by people to find information and make sense of it and also by computer programs to reason about data. There are many types of KOS, such as glossaries, thesauri, classifications, and ontologies, among others. Although in some cases the term *ontology* is loosely used as a synonym of KOS (e.g. McGuinness, 2003), an ontology is often defined in knowledge engineering and artificial intelligence as 'a specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects' (Gruber, 1993, p. 199). As stated by Komninos et al. (2015, p. 34), 'the use of ontologies in the field of smart cities is a relatively new field of research'. Exploring the scholarly literature in the last five years revealed that the most significant smart-city sectors for which ontologies have been developed are crisis management, economics, education, energy, environment, health, home, social welfare, and urban planning.

Ontologies for crisis management cover activities related to hazardous situations and disasters. Elmhadhbi et al. (2019) described POLARISCO (Plateforme OpérationnelLe d'Actualisation du Renseignement Interservices pour la Sécurité Civile Ontology), a modular ontology that enables semantically interoperable communication among emergency responders in large-scale crisis situations. Similarly, Gaur et al. (2019) introduced the *empathi* ontology, which conceptualises core concepts in the domain of emergency management. Kurte et al. (2019) presented a semantics-based approach for modelling spatio-temporal changes using a series of multi-temporal remote sensing images acquired during a flood disaster. In particular, a semantic model called Dynamic Flood Ontology was used to represent spatio-temporal information. Chehade et al. (2020) developed ResOnt, an ontology that aims to support situation awareness and communication in rescue operations by dealing with all components, aspects, and factors relevant in such situations.

In the economics sector, Cantone et al. (2020) described an extension of OASIS (Ontology for Agents, Systems, and Integration of Services) (Cantone et al., 2019), an ontology that defines a request-execute communication protocol for the Internet of Agents (IoA). In this case, OASIS was extended to include contract terms that help construct ontological representations of smart contracts secured on the blockchain. Such terms also include conditionals applied to constrain agent actions or trigger them when suitable conditions hold.

Most ontologies exploring information technology in education are related to eLearning. Bouzidi et al. (2019) proposed OntoGamif (Ontology of Gamification), which provides a semantic model that describes concepts in the gamification domain. OntoGamif, organised as seven interlinked gamification-related modules (i.e. core concepts, organisation, psychology, evaluation, ethics, risks, and the user), was designed to support not only the work of gamification system developers but also the

mutual understanding between gamification service providers and consumers. Chimalakonda & Nori (2020) proposed an ontology-based framework for modelling different aspects of instructional design. In particular, three modular ontologies, i.e. an ontology for goals, an ontology for instructional processes and an ontology for instructional materials, can be employed to represent the instructional design.

In the energy sector, Kott & Kott (2019) proposed an ontology for effectively managing energy in household devices and for the reduction of greenhouse gases. The ontology, where the two largest groups of ontological categories refer to electrical equipment (i.e. home appliances and energy production systems), is intended to present end users as prosumers that can change their energy consumption profiles. Spoladore et al. (2019) introduced ComfOnt, which aimed to represent features of the Smart Home ecosystem such as energy consumption, household devices (i.e. appliances, sensors, and actuators), and indoor comfort. The ontology comprises four primary modules that interact with each other: a Person uses Devices that can contribute to implementing Comfort metrics within a Domestic Environment. Chun et al. (2020) proposed the Energy Knowledge Graph as an upper ontology for integrating knowledge resources in energy systems. To this end, they employed the IoT Architectural Reference Model (Bassi et al., 2016) as the top-level ontology, specifying concepts at the highest abstraction level to facilitate a common understanding of the IoT domain.

Most environmental ontologies deal with air, water and soil quality monitoring and management. In this context, Hu et al. (2020) described the construction of an ontology for multi-sensor information integration in managing urban gardens and green spaces.

Ontologies in the healthcare sector primarily focus on providing solutions to increase the performance of medical services. Peng & Goswami (2019) presented a method that applies Semantic Web technologies to integrate health data and home environment data supplied by heterogeneous services and devices into a resource graph. In particular, the Linked Health Resources framework was developed to integrate the most popular health-related data sources. Tiwari & Abraham (2020) presented the Smart Health-Care Ontology (SHCO), designed to deal with healthcare information and IoT devices. This study also proposed a methodology to assess the semantic model of SHCO based on different test cases, apart from analysing the quality of the knowledge with various verification and validation tools. Moreira et al. (2020) described the SAREF4health extension, the first attempt to extend the SAREF (Smart Applications REFerence) ontology for specific e-Health use cases. In particular, SAREF4health was developed from an ontology-driven conceptual modelling approach, where an electrocardiography ontology grounded in the Unified Foundational Ontology played the role of a reference model.

Ontologies for home automation aim to benefit not only the individuals living in the house but also the whole of society and the external environment. Balaji et al. (2018) described Brick, a complete and expressive metadata schema for representing sensors, subsystems and the relationships between them in energy-efficient buildings so that smart-building applications can be developed.

Most ontologies about social welfare mainly focus on the issue of poverty. Nasim & Khan (2018) presented an ontological approach to model the problem of poverty, describing the different ways through which the issue of poverty can be addressed. Moreover, the reasoning capability of the ontology revealed the relationships between various causes of poverty at different levels of granularity. Panori et al. (2019) proposed an ontology structure for measuring urban poverty through smart-city applications, where three levels of input data can be defined: spatio-temporal information, basic demographic information of individuals, and essential information for computing the Multidimensional Poverty Index (MPI).

Finally, ontologies developed in the urban-planning sector support decision-making on infrastructures and services to citizens. Hellmund et al. (2018) introduced the HERACLES (HEritage Resilience Against CLimate Events on Site) ontology for preserving cultural heritage considering the

effects of climate change, where Action mitigates Damage and Effect caused by climate factors that occurred on a Cultural Heritage Element. Wei et al. (2020) presented a web-based decision support system for urban infrastructure inter-asset management. This system consists of a suite of interlinked modular ontologies that model the knowledge on infrastructure assets (e.g. road, ground, cable), triggers (e.g. pipe leaking), consequences (e.g. traffic disruption), and investigation techniques (e.g. sensors) for urban infrastructure management. Viktorović et al. (2020) proposed the Connected Traffic Data Ontology (CTDO), which aims to understand the state of traffic within urban environments at any given moment. This ontology, compatible with the SOSA (Sensor, Observation, Sample, and Actuator) ontology (Janowicz et al., 2018), provides a more suitable model for large volumes of time-sensitive data coming from multi-sensory platforms, where the use cases of connected autonomous vehicles were considered.

It can be concluded that ontology development for the Smart City framework is currently an active field of research. Most of these ontological initiatives share two characteristics. On the one hand, most smart-city ontologies are aimed at modelling highly restricted domains of interest, e.g. Smart Bike Sharing (Syzdykbayev et al., 2019), Smart Parking, Smart Garbage Control, Smart Streetlight, and Smart Complaint management (Qamar et al., 2019), and others shown in the above studies. In contrast, Baracho et al. (2019) and Soergel et al. (2020) presented a proposal for developing a comprehensive organisation system for smart cities, smart buildings and smart life, where the last aims to improve daily activities, e.g. keeping track of the contents of a refrigerator; however, their proposal was very preliminary. On the other hand, most smart-city ontologies have been designed and developed for physical-sensor networks. For example, Espinoza-Arias et al. (2019) performed a survey of fifteen smart-city ontologies, where they analysed features about the development process, the availability, and the type of domain. The study revealed that most of the reviewed ontologies focused on sensors and the Internet of Things (IoT). It should be noted that there have been some research efforts to develop semantic models that could deal with messages reported through digital participatory platforms. For example, city departments in the USA provide stakeholders (e.g. citizens, corporations, etc.) with two types of services: emergency services accessed by 911 and non-emergency services accessed by 311. In this context, Nalchigar & Fox (2018) developed the Open311 ontology, which was designed to create a common vocabulary that could support interoperability among Open311 datasets. The specific topics on which this ontology focuses are found in the class SUBJECT, which subsumes the classes GARBAGECONTAINER, PEST, PLANTS, PROPERTY, ROADSYMBOL, and TRANSPORTATIONROUTES, each one further decomposed into other subclasses. The study demonstrated that this ontology could contribute to mapping the diverse content of existing Open311 datasets into a single semantic model. However, it was not designed to support the processing of citizens' text messages. Rani et al. (2016) employed the Open311 ontology to automatically annotate keywords in citizens' queries, but this study has several limitations. In particular, the ontology was implemented for a highly constrained domain and tested with an ad-hoc dataset, as reflected by the fact that the system only focused on eight classes (i.e. GRASS, INSECTPEST, INTERNET, NOISE, SMOKING, STREETLIGHT, TREE, and WASTE) and the NLP module relied on a simplistic keyword-recognition procedure based on lists of synonyms. In contrast, the DIAPASON ontology, which supports topic categorisation and keyword recognition with UGC, can model a large variety of social and physical community-problem types from the perspective of citizens living in a Smart Society.

3. The DIAPASON ontology

3.1. Methodological approach

The definition of a solid methodology serves as a framework for designing and developing a

consensual ontology within and between research teams, defining the activities that should be performed and the order of execution of such activities. In this regard, the construction of the DIAPASON ontology is based on METHONTOLOGY (Fernández López et al., 1997, 1999) guidelines for conducting the entire development process. Moreover, we adopted the ontology-engineering paradigm of ODPs (Hitzler et al., 2016), particularly Krisnadhi and Hitzler's (2016) approach to developing ODPs effectively. In this way, we facilitated the development of a modularised ontology and contributed to sharing our ontology in the form of new and sufficiently general ODPs. In this regard, the DIAPASON ontology resulted from integrating the modules corresponding to two ODPs: the structural ODP that takes the form of a subClassOf hierarchy and the content ODP that conceptualises the semantics of specific problem types. As shown in Fig. 2, the ontology development process was structured into two main stages, i.e. one centred around the ODPs and the other around problem schemas.

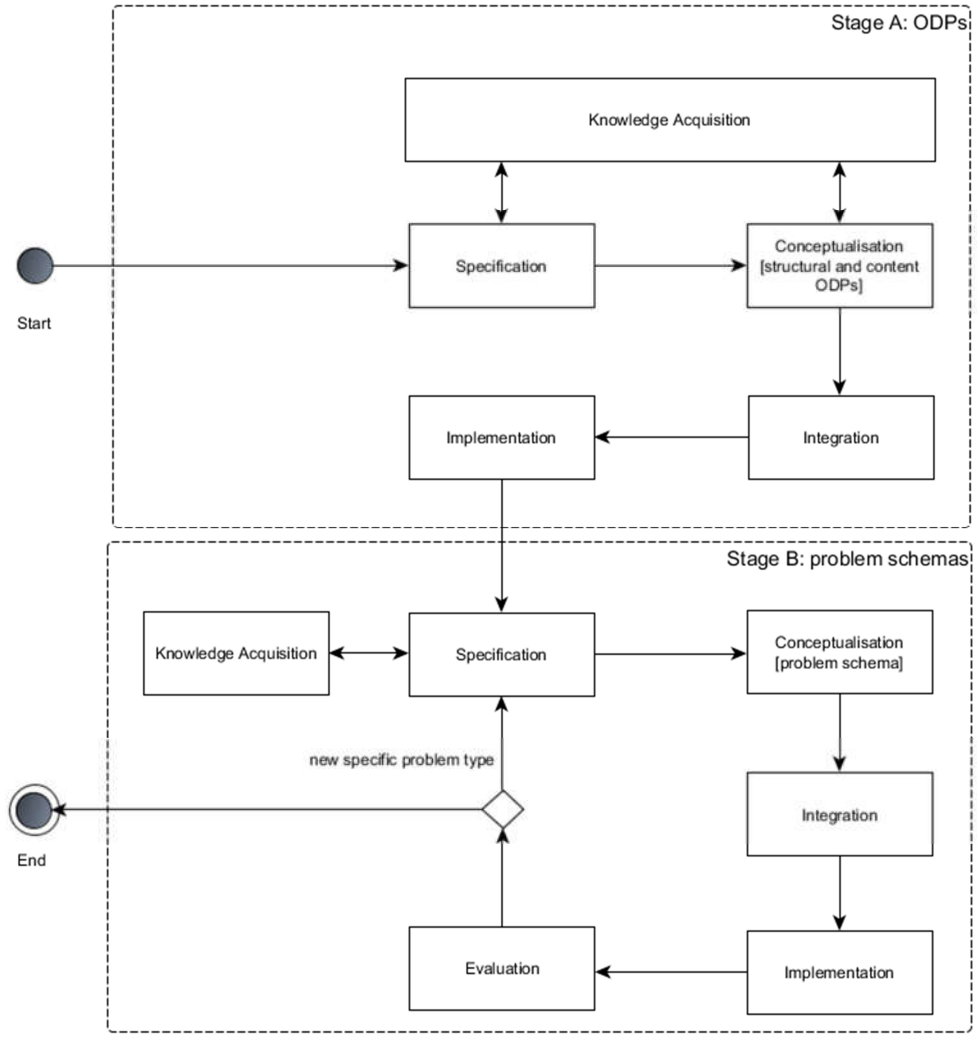

Fig. 2. Constructing the DIAPASON ontology.

3.2. Stage A: Specification and Knowledge Acquisition

The Specification activity at Stage A determines the purpose and scope of the ontology, its intended use, and target users. The purpose of the DIAPASON ontology is to assist the NLP module in the tasks of topic categorisation and keyword recognition when analysing English and Spanish short texts from social media and digital participatory platforms, as the system aims to automatically recognise a wide variety of community problems in all the core domains relevant to the development of smart societies for entrepreneurial, political or social actions. For experimental purposes, the system will result in an online workbench where researchers could explore their UGC corpora to gain situational awareness.

Moreover, a list of competency questions was elaborated, which the system is expected to answer over the data in the ontology. In other words, competency questions aim to express the scope or functionality of the system, so they represent the ontology requirements. Considering that DIAPASON leverages social sensors to report to citizens and policymakers about the problems that hinder the development of smart societies, the list of competency questions is as follows:

a) Which types of problems in [realm/dimension/domain] were detected in [instant/interval]* in [location]*?
b) Find text messages related to [realm/dimension/domain/problem type]* in [instant/interval]* in [location]*.
c) Find all the dimensions/domains/problem types associated with [keyword]* in [language]* in [instant/interval]* in [location]*.
d) Find all the keywords in [language] related to [realm/dimension/domain/problem type]* in [instant/interval]* in [location]*.
e) When/where was [problem type] detected?

Indeed, the above list is presented as an inventory of patterns on competency questions, where components marked with an asterisk are optional, and elements such as realm, dimension, domain, and problem type are described in Section 3.3.1. Therefore, each pattern can be instantiated in one or more competency questions that take different forms depending on the problem types analysed and the UGC corpus explored. For example, suppose that we compile a collection of tweets dated 2021 with the hashtag #brightonbeach. After populating the DIAPASON ontology with the data extracted by the text-processing module, some of the queries for this input dataset could be as follows:

i) Which types of problems in the ecological-hazard domain were detected between September 2021 and December 2021?
ii) Find text messages related to ecological hazards.
iii) Find all the problem types associated with the keyword *sewage*.
iv) Find all the keywords in Spanish related to the environmental dimension.
v) When were seawater-quality problems detected?

As shown above, competency questions were formulated according to the expected purpose of the ontology, that is, detecting problem themes of different granularity (i.e. *problem realm*, *problem dimension*, *problem domain*, and *problem type*) recognised at a given *time* and *location* and summarised through *keywords* revealed from the analysis of *text messages* written in a given *language*. Finally, core elements for the ontology were identified by considering the inventory of patterns on competency questions, the type of content that should be analysed, and the purpose of the analysis in DIAPASON.

DIAPASON explores community problems in the context of situational awareness, which refers to 'knowing what is going on around you' in a decision-making process (Endsley, 2000, p. 5). According

to Endsley (1995), the three primary components of situation awareness in a given environment are (a) the perception of the elements in the environment, (b) the comprehension of the current situation, and (c) the projection of future actions. In this regard, lexical items in UGC describe the situational context (i.e. objects and people in one's surroundings), from which topic categorisation and keyword recognition reveal what people are thinking and doing (i.e. the comprehension of the current situation).

DIAPASON also adopts the subjectivist approach to viewing community problems. It should be noted that the sociological definition of social problems has been conventionally studied from two competing perspectives, i.e. objectivism and subjectivism. Our common-sense notion of what social problems are brings to mind some characteristics of the conditions in which they occur (Loseke, 2017):

a) The conditions cause harm. In the context of our study, such conditions disrupt QoL and undermine well-being.
b) The conditions are believed to affect a significant number of people. In our case, social problems are not personal troubles but problems that affect the community.
c) The conditions can be changed, as social problems can be fixed.
d) The conditions should be changed, so we expect something to be done to fix the problem. For example, citizens can employ digital participatory platforms to report non-emergency problems about urban infrastructure and services so that relevant authorities can take corrective measures.

Therefore, this definition of social problems is presented in terms of conditions that are troublesome, prevalent, changeable, and in need of change. Social problems are viewed as objective conditions, focusing on the real physical world.

However, occurrences of such conditions do not necessarily imply the existence of social problems, as illustrated by sexual discrimination and global warming (Best, 1995). Nowadays, sexism is widely understood to be a social problem; however, sexism was not perceived as a form of discrimination in the USA before 1970, when the feminist movement began to gain public attention. Something similar may also occur even when harmful conditions seem purely objective. For example, although scientists have demonstrated that certain human activities increase greenhouse gases and thus affect extreme weather events and health risks, climate-change sceptics do not regard global warming as a social problem. These examples show that identifying what is or is not a social problem cannot be grounded on objective conditions (e.g. treating women less favourably because of their sex or increasing the planet's overall temperature) but on subjective judgments, which are subject to change. Best (1995, p. 4-5) concluded that 'social problems are what people view as social problems [...] No condition is a social problem until someone considers it a social problem', which represents the subjectivist approach to social problems:

> [...] it is not an objective quality of a social condition, but rather the subjective reactions to that condition, that make something a social problem. Therefore, social problems should not be viewed as a type of social condition, but as a *process* of responding to social conditions. (Best, 2017, p. 7-8)

Although subjectivism was adopted in sociology as a perspective to analyse social problems, i.e. community problems in the social realm of the DIAPASON ontology, the same approach can be adopted with community problems in the physical realm. For example, fast driving over large potholes can cause severe damage to cars and make drivers lose control of their vehicles. Suppose a road with a pothole. Although a troubling condition exists, if citizens do not complain about it, there will be no public worry and thus no community problem. Therefore, we can conclude that the pothole is not a problem by itself. The problem arises only when people make claims about it.

Community problems should be understood in terms of a process that subjectivists call "claimsmaking", which is the only thing all these problems have in common. Claimsmaking 'determines both which phenomena will be designated as social problems, and which characteristics are attributed to those problems' (Letukas, 2014, p. 49). In this context, Best (2017) identified three elements that play a key role in this process: (a) claims, i.e. statements about social problems that need to be solved, (b) claimsmakers, i.e. the people that make claims, and (c) troubling conditions, i.e. the conditions that become the subject of the claims. In claimsmaking, therefore, claimsmakers argue that a particular troubling condition is related to a given social problem (i.e. community problem) through a claim that brings the condition to the attention of the audience. In turn, claims are structured into three components: grounds (i.e. information about the troubling conditions), warrants (i.e. justifications for taking action), and conclusions (i.e. recommended actions) (Best, 1990). Moreover, typification, i.e. characterising problems' nature, is crucial in exploring social problems. In our study, claimsmakers are citizens, claims take the form of UGC, and the analysis of the grounds of claims is oriented to problem detection, for which problem schemas include typified knowledge.

At this stage, Knowledge Acquisition was carried out to support activities such as Specification and Conceptualisation. Relevant knowledge sources were not only the Smart City frameworks in Section 2.1 but also QoL reports, encyclopedias and textbooks, where problems are explored from a global perspective, and research articles and other scholarly publications, where particular aspects of problems are examined in depth. At the next stage, we designed the DIAPASON ontology based on the structural ODP, which provides the taxonomic hierarchy, and the content ODP, which conceptualises the semantics of specific problem types. Whereas the former is presented in Section 3.3, the latter is addressed in Section 3.4.

3.3. Stage A: Conceptualisation of the structural ODP

3.3.1. Overview

As shown in Fig. 3, the taxonomic hierarchy of the DIAPASON ontology consists of four levels of thematic granularity, each one including a different kind of element: problem realm, problem dimension, problem domain, and specific problem type.

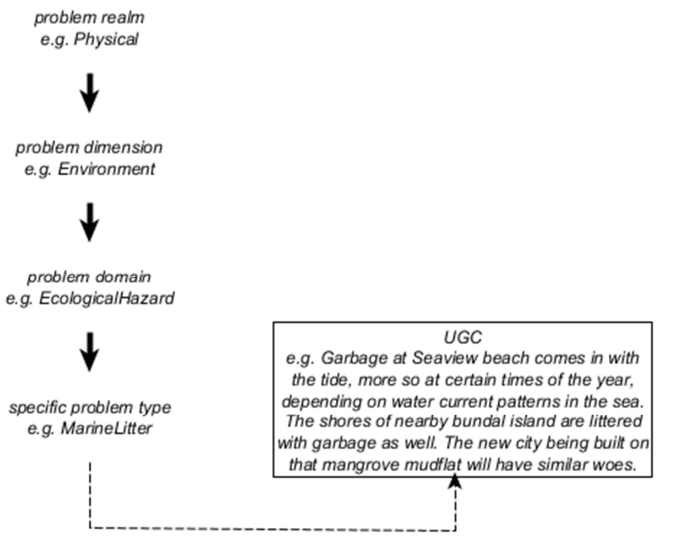

Fig. 3. Ontological levels in DIAPASON.

The upper level is structured into two problem realms: the social realm, which subsumes the LIVING, ECONOMY and GOVERNANCE problem dimensions, and the physical realm, which subsumes the MOBILITY, INFRASTRUCTURE and ENVIRONMENT problem dimensions. The modelling of the dimension level primarily results from Smart City frameworks (Giffinger et al., 2007; Govada et al., 2017; Appio et al., 2019). At the next level, each problem dimension is organised in problem domains on which citizens can show an attitude of disapproval towards some specific aspect of the community. QoL projects primarily influenced the modelling of the domain level. Finally, the lower level describes specific types of community problems that can affect some, most or all citizens for each problem domain.

Problem realms, problem dimensions, problem domains and specific problem types were modelled in a single class hierarchy, being PROBLEM the superclass at the top. In other words, taking Fig. 3 as an example, the class PHYSICAL should be interpreted as conceptually representing the problems about the physical realm, the class ENVIRONMENT as the problems about the environmental dimension, the class ECOLOGICALHAZARD as the problems about the ecological-hazard domain, and the class MARINELITTER as the problems about marine litter. This ontological decision was guided by the fact that the DIAPASON ontology is oriented to topic categorisation. Figure 4 shows the structural ODP (i.e. the subClassOf hierarchy) containing realms, dimensions and domains, as these levels are fully developed. The level of specific problem types is currently under development because of the numerous elements that can be found. Each specific problem type is modelled as a distinct class subsumed by one or more classes at the domain level. For example, MARINELITTER is a subclass of ECOLOGICALHAZARD, and DISAPPOINTMENTGOVERNMENT is a subclass of POLITICALACTIVITY.

3.3.2. Problem dimensions and domains

The problem domains in the LIVING dimension were primarily determined by considering QoL reports (e.g. Eurostat, 2017; Bundesregierung, 2018; Office for National Statistics, 2019; Helliwell et al., 2020; Istituto Nazionale di Statistica, 2020; OECD, 2020). Sociological encyclopedias and textbooks (e.g. Parrillo, 2008; Eitzen et al., 2014; Marginean, 2014; Seccombe & Kornblum, 2020) were also examined to explore this dimension. In the end, eleven domains were identified:

a) CRIMEJUSTICE deals with the integrity of the individual, as the measure of right and the response to wrong, particularly with security issues such as criminality, violence, and vandalism.
b) CULTURE encompasses aspects such as language, literature, music and song, non-verbal communication, religion or belief systems, rites and ceremonies, sport and games, food, clothing, customs and traditions, and the arts, among others, through which individuals, groups of individuals, and communities express their humanity and the meaning they give to their existence.
c) EDUCATIONTRAINING focuses on the knowledge, competences, skills and activities related to formal education and lifelong learning.
d) HEALTH refers to three main areas: (a) health status, i.e. physical and mental problems of the individual, (b) risk factors for the individual's health, resulting from lifestyle or personal choices (e.g. obesity, dietary habits, tobacco, and alcohol consumption), and (c) access to healthcare services, facing barriers such as cost, distance, and waiting times, among others.
e) INEQUALITY includes, but is not limited to, three primary types of discrimination against a group in a society, i.e. racial and ethnic, gender, and economic inequalities.
f) LEISURE refers to the free time outside work and the sporting activities and recreation that people undertake during that time.

g) POPULATION is a broad category that primarily groups together aspects reflecting cumulative processes on population (e.g. life expectancy and increase/decrease of the population) and issues related to specific segments of the population (e.g. the disabled and the elderly).
h) RESEARCHINNOVATION focuses on any issue related to the creation, application or diffusion of knowledge (e.g. patent protection and brain drain).
i) SOCIALRELATIONSHIP is based on two main areas: personal relationships (e.g. family and friends) and community relationships, i.e. civic engagement (e.g. volunteering).
j) SUBJECTIVEWELLBEING represents a personal exploration of how the individual feels. In particular, the individual evaluates their life as a whole in terms of dissatisfaction, which requires some reflection on multiple aspects of past experiences regarding personal standards (i.e. overall assessment of life), or the individual indicates negative emotions about their current situation (i.e. affective state).
k) TERRORISMWAR involves extreme violence motivated by political, ideological or strategic aims and inflicted by one group of individuals against another.

The ECONOMY dimension corresponds to material well-being, which is about satisfaction with a variety of economic concerns: personal and family financial situation, perceived income adequacy, material possessions and the extent to which physical needs are met (i.e. standard of living), feelings of financial security, feelings about access to major goods and services, aspirations and attainment of material goods, and marketplace activities (Sirgy, 2018). In this regard, the ECONOMY dimension comprises five main domains: CONSUMPTION (e.g. personal expenditure), INCOME (e.g. purchasing power), MACROECONOMICS (e.g. inflation, investment rate, national debt, and public expenditure), WEALTH (e.g. financial assets), and WORK (e.g. stability, over-qualification, unemployment, and work-life balance).

The GOVERNANCE dimension, identified with politics, was grounded on Christensen's model (2016). From a sociological perspective, this author explored the interpretation of political dissatisfaction based on two issues: political support, which is an issue of satisfaction with democracy and trust in institutions, politicians and political parties, and subjective political empowerment, which concerns the extent to which citizens feel they can influence political decisions. In DIAPASON, the entity that raises political dissatisfaction was the focus of attention. Therefore, the GOVERNANCE dimension comprises five domains: INSTITUTIONS, POLITICALACTIVITY, POLITICALPARTIES, POLITICALSYSTEM, and POLITICIANS.

The MOBILITY dimension refers to the capacity of a given population to move from one place to another, where differences lie in the mode of travel (Böhler-Baedeker & Durant, 2015). From this context, a functional perspective of the definition of *traffic* was employed, thus referring to 'activities designed to move objects and people around, with the help of means of transportation and the corresponding infrastructure' (Götz, 2014, p. 6705). Therefore, MOBILITY is modelled as a dimension that is structured into different types of traffic, each one corresponding to a distinct domain: MOTORIZEDVEHICULARTRAFFIC (e.g. buses, cars, e-scooters, motorcycles, taxis, and trucks), NONMOTORIZEDVEHICULARTRAFFIC (e.g. bicycles, skateboards, and roller-skates), and PEDESTRIANTRAFFIC.

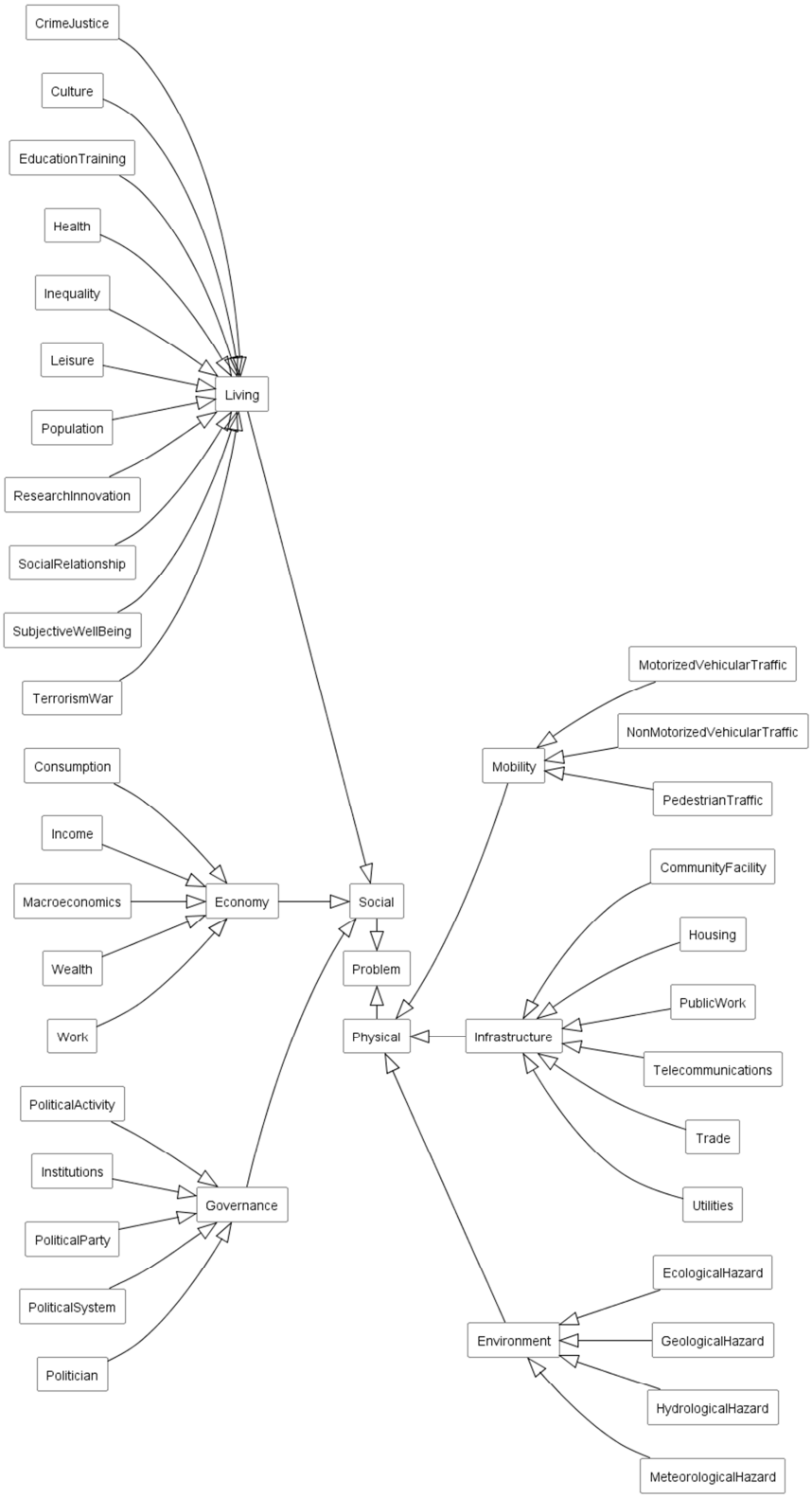

Fig. 4. Realms, dimensions and domains in the DIAPASON ontology.

The INFRASTRUCTURE dimension includes the built facilities and services that give cities their form and function. The INFRASTRUCTURE dimension was structured into six domains (Howes & Robinson, 2005; Neuman, 2005; Parrillo, 2008): COMMUNITYFACILITY (e.g. schools, parks, hospitals, libraries, prisons, emergency services, and civic buildings), HOUSING (e.g. structural problems of the dwelling and overcrowding), PUBLICWORK (e.g. roads and bridges, dams and canals, ports and airports, and railways), TELECOMMUNICATIONS (e.g. mobile, radio, television, and the Internet), TRADE (e.g. factories, offices, shops, and warehouses), and UTILITIES (e.g. gas and electricity, water supply, sewerage, and waste collection and disposal).

The ENVIRONMENT dimension is about how the environment can affect our QoL and how our activities can affect the environment. The domains in the ENVIRONMENT dimension were organised according to the common types of hazards identified in books such as Lindell et al. (2006), Schwab et al. (2007), and Haddow et al. (2011): ECOLOGICALHAZARD (e.g. oil spills, pollution, and wildfires), GEOLOGICALHAZARD (e.g. earthquakes, landslides, and volcanic eruptions), HYDROLOGICALHAZARD (e.g. drought and floods), and METEOROLOGICALHAZARD (e.g. hail, heat waves, and tornadoes).

3.4. Stage A: Conceptualisation of the content ODP and Integration

From the core elements identified in the Specification activity, a content ODP was created to describe the semantics of specific problem types with respect to the UGC through which such problems are reported. First, the content ODP was modelled with the corresponding classes, object properties and datatype properties, graphically represented in Fig. 5. Second, the T-Box (classes) and R-Box (relations) axiomatisation of the content ODP was provided in Description Logics (DL):[1]

%Comment: 'Classes: subClassOf hierarchy'.

(a1) PROBLEM ⊑ CCO:INFORMATIONCONTENTETITY

(a2) PROBLEMCONTENT ⊑ CCO:INFORMATIONCONTENTENTITY

(a3) PROBLEMREPRESENTATION ⊑ MF:COGNITIVEREPRESENTATION

(a4) PROBLEMBELIEF ⊑ MF:BELIEF

(a5) MF:MENTALPROCESS ⊑ MF:BODILYPROCESS

(a6) USERGENERATEDCONTENTUNIT ⊑ CCO:INFORMATIONBEARINGARTIFACT

(a7) USERGENERATEDCONTENTITEM ⊑ CCO:INFORMATIONBEARINGARTIFACT

(a8) VOICEMESSAGE ⊑ USERGENERATEDCONTENTITEM

(a9) TEXTMESSAGE ⊑ USERGENERATEDCONTENTITEM

(a10) CCO:IMAGE ⊑ USERGENERATEDCONTENTITEM

(a11) CCO:VIDEO ⊑ USERGENERATEDCONTENTITEM

(a12) USERGENERATEDCONTENT ⊑ CCO:INFORMATIONCONTENTENTITY

(a13) DESCRIPTOR ⊑ CCO:DESCRIPTIVEINFORMATIONCONTENTENTITY

(a14) KEYWORD ⊑ CCO:DESCRIPTIVEINFORMATIONCONTENTENTITY

(a15) SYNSET ⊑ CCO:NONNAMEIDENTIFIER

(a16) BFO:SPATIALREGION ⊑ BFO:IMMATERIALENTITY

(a17) BFO:TEMPORALINSTANT ⊑ BFO:TEMPORALREGION

[1] Krötzsch et al. (2014) provided an introduction to DL, where the main concepts and features were explained with examples.

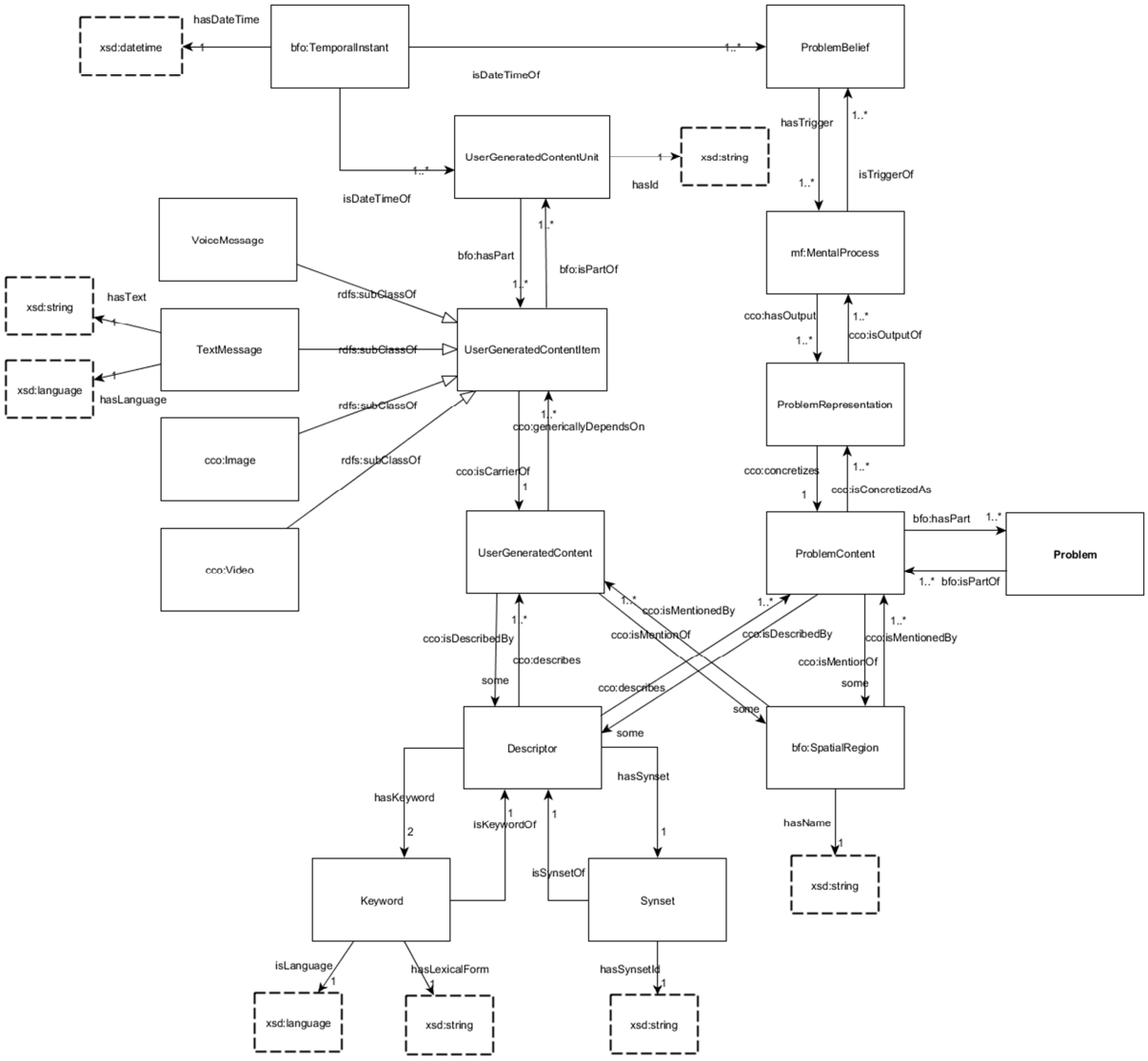

Fig. 5. The problem-type ODP.

%Comment: 'Classes: disjointness'.

(a18) DESCRIPTOR ⊑ ¬KEYWORD

(a19) USERGENERATEDCONTENTUNIT ⊑ ¬USERGENERATEDCONTENTITEM

%Comment: 'Object properties: domain, range and cardinality restriction'.

(a20) PROBLEM ⊑ ≥1 bfo:isPartOf.PROBLEMCONTENT

(a21) PROBLEMCONTENT ⊑ (∃cco:isDescribedBy.DESCRIPTOR) ⊓ (∃cco:isMentionOf.BFO:SPATIALREGION)

(a22) PROBLEMCONTENT ⊑ (≥1 bfo:hasPart.PROBLEM) ⊓ (≥1cco:isConcretizedAs.PROBLEMREPRESENTATION)

(a23) PROBLEMREPRESENTATION ⊑ (≥1 cco:concretizes.PROBLEMCONTENT) ⊓ (≥1 cco:isOutputOf.MF:MENTALPROCESS)
(a24) MF:MENTALPROCESS ⊑ (≥1 cco:hasOutput.PROBLEMREPRESENTATION) ⊓ (≥1 isTriggerOf.PROBLEMBELIEF)
(a25) PROBLEMBELIEF ⊑ ≥1 hasTrigger.MF:MENTALPROCESS
(a26) BFO:TEMPORALINSTANT ⊑ (≥1 isDateTimeOf.PROBLEMBELIEF) ⊓ (≥1 isDateTimeOf.USERGENERATEDCONTENTUNIT)
(a27) USERGENERATEDCONTENTUNIT ⊑ ≥1 bfo:hasPart.USERGENERATEDCONTENTITEM
(a28) USERGENERATEDCONTENTITEM ⊑ (≥1 bfo:isPartOf.USERGENERATEDCONTENTUNIT) ⊓ (=1 cco:isCarrierOf.USERGENERATEDCONTENT)
(a29) USERGENERATEDCONTENT ⊑ ∃cco:isMentionOf.BFO:SPATIALREGION ⊓ ∃cco:isDescribedBy.DESCRIPTOR
(a30) USERGENERATEDCONTENT ⊑ ≥1 cco:genericallyDependsOn.USERGENERATEDCONTENTITEM
(a31) BFO:SPATIALREGION ⊑ (≥1 cco:isMentionedBy.USERGENERATEDCONTENT) ⊓ (≥1 cco:isMentionedBy.PROBLEMCONTENT)
(a32) DESCRIPTOR ⊑ (≥1 cco:describes.USERGENERATEDCONTENT) ⊓ (≥1 cco:describes.PROBLEMCONTENT) ⊓ (=2 hasKeyword.KEYWORDS) ⊓ (=1 hasSynset.SYNSET)
(a33) KEYWORD ⊑ =1 isKeywordOf.DESCRIPTOR
(a34) SYNSET ⊑ =1 isSynsetOf.DESCRIPTOR

%Comment: 'Object properties: inverse properties'.
(a35) cco:describes ⊆ cco:isDescribedBy⁻
(a36) bfo:isPartOf ⊆ bfo:hasPart⁻
(a37) cco:concretizes ⊆ cco:isConcretizedAs⁻
(a38) cco:isOutputOf ⊆ cco:hasOutput⁻
(a39) cco:isTriggerOf ⊆ cco:hasTrigger⁻
(a40) cco:isCarrierOf ⊆ cco:genericallyDependsOn⁻
(a41) cco:isMentionOf ⊆ cco:isMentionedBy⁻
(a42) isKeywordOf ⊆ hasKeyword⁻
(a43) isSynsetOf ⊆ hasSynset⁻
(a44) cco:isDescribedBy ⊆ cco:describes⁻
(a45) bfo:hasPart ⊆ bfo:isPartOf⁻
(a46) cco:isConcretizedAs ⊆ cco:concretizes⁻
(a47) cco:hasOutput ⊆ cco:isOutputOf⁻
(a48) cco:hasTrigger ⊆ cco:isTriggerOf⁻
(a49) cco:genericallyDependsOn ⊆ cco:isCarrierOf⁻
(a50) cco:isMentionedBy ⊆ cco:isMentionOf⁻
(a51) hasKeyword ⊆ isKeywordOf⁻
(a52) hasSynset ⊆ isSynsetOf⁻

%Comment: 'Datatype properties: domain, range and cardinality restriction'.
(a53) USERGENERATEDCONTENTUNIT ⊑ =1 hasId.xsd:string
(a54) TEXTMESSAGE ⊑ (=1 hasText.xsd:string) ⊓ (=1 hasLanguage.xsd:language)

(a55) KEYWORD ⊑ (=1 hasLexicalForm.xsd:string) ⊓ (=1 isLanguage.xsd:language)
(a56) SYNSET ⊑ =1 hasSynsetId.xsd:string
(a57) BFO:SPATIALREGION ⊑ =1 hasName.xsd:string
(a58) BFO:TEMPORALINSTANT ⊑ =1 hasDateTime.xsd:datetime

This DL axiomatisation, whose expressivity is ALCHIQ(D), is compliant with the OWL2-DL profile, one of the most expressive members of the OWL family, thus contributing to making the ontology interoperable.

Ontological decisions were guided by the fact that the DIAPASON ontology is oriented to topic categorisation (PROBLEM) and keyword recognition (KEYWORD). A critical issue for understanding the choice of the classes and properties in this content ODP is the architecture of ontology development in DIAPASON. As shown in the subClassOf Hierarchy in Fig. 6, a multi-tiered approach to ontology development was adopted by integrating the specialised ontology into top- and mid-level ontologies to facilitate semantic interoperability. At the lower level, each root class in the structural and content ODPs (Fig. 4 and Fig. 5) that construct the specialised-domain ontology was connected to an existing class in one of the modules in the Common Core Ontologies framework (CCO)[2] or Mental Functioning Ontology (MF),[3] which were employed as mid-level ontologies, or in the Basic Formal Ontology (BFO),[4] which was employed as the top-level ontology. Finally, each root class in the mid-level ontology was subsumed by a BFO class. Therefore, the DIAPASON ontology is BFO-compliant. The top- and mid-level ontologies contribute to integrating information and sharing knowledge among heterogeneous sources.

On the one hand, problem-related classes (i.e. PROBLEM, PROBLEMCONTENT, PROBLEMREPRESENTATION, and PROBLEMBELIEF) are of paramount importance for DIAPASON, where MF played a central role in their modelling. The rationale for having such problem-related classes is grounded on our subjectivist approach to community problems. As explained in Section 3.2, identifying what is or is not a problem cannot rely on objective conditions in the real physical world but on subjective judgments about such conditions, providing that people negatively react to such conditions through claimsmaking. In this regard, such judgments or thoughts represent citizens' beliefs about some portion of reality. For example, in saying that I believe that carbon dioxide emissions from transport increase health risks, I am saying something about how I view the world. It should be noted that problem beliefs result from personal perceptions and judgments, so they cannot be interpreted as objective truths but as plausible truths that can be transformed into knowledge. Therefore, problem beliefs are taken by the individual to be true, although they may or may not actually be true. The next issue is to determine which type of belief can represent community problems.

In philosophy, there is the traditional position of regarding beliefs as having a dual nature. Whereas a dispositional belief is information available to mind for endorsement, an occurrent belief is a thought consciously endorsed (Rose & Schaffer, 2013). In other words, a dispositional belief is a disposition that can be realised in an occurrent belief at a particular time, which involves the process of bringing belief to the conscious (i.e. the forefront of our minds). For example, Peter may believe, even when he is not thinking about it (e.g. sleeping), that carbon dioxide emissions from transport increase health risks. This dispositional belief is not being continuously activated in his mind but persists through time, so it is categorised as a continuant. However, when he tweeted that '20mph roads actually increase air pollution as engines cause massive tailbacks which add to the toxic soup', his dispositional belief was deliberately manifested in an occurrent belief by the conscious process of thinking at that moment that

[2] https://github.com/CommonCoreOntology/CommonCoreOntologies
[3] https://github.com/jannahastings/mental-functioning-ontology
[4] https://basic-formal-ontology.org/

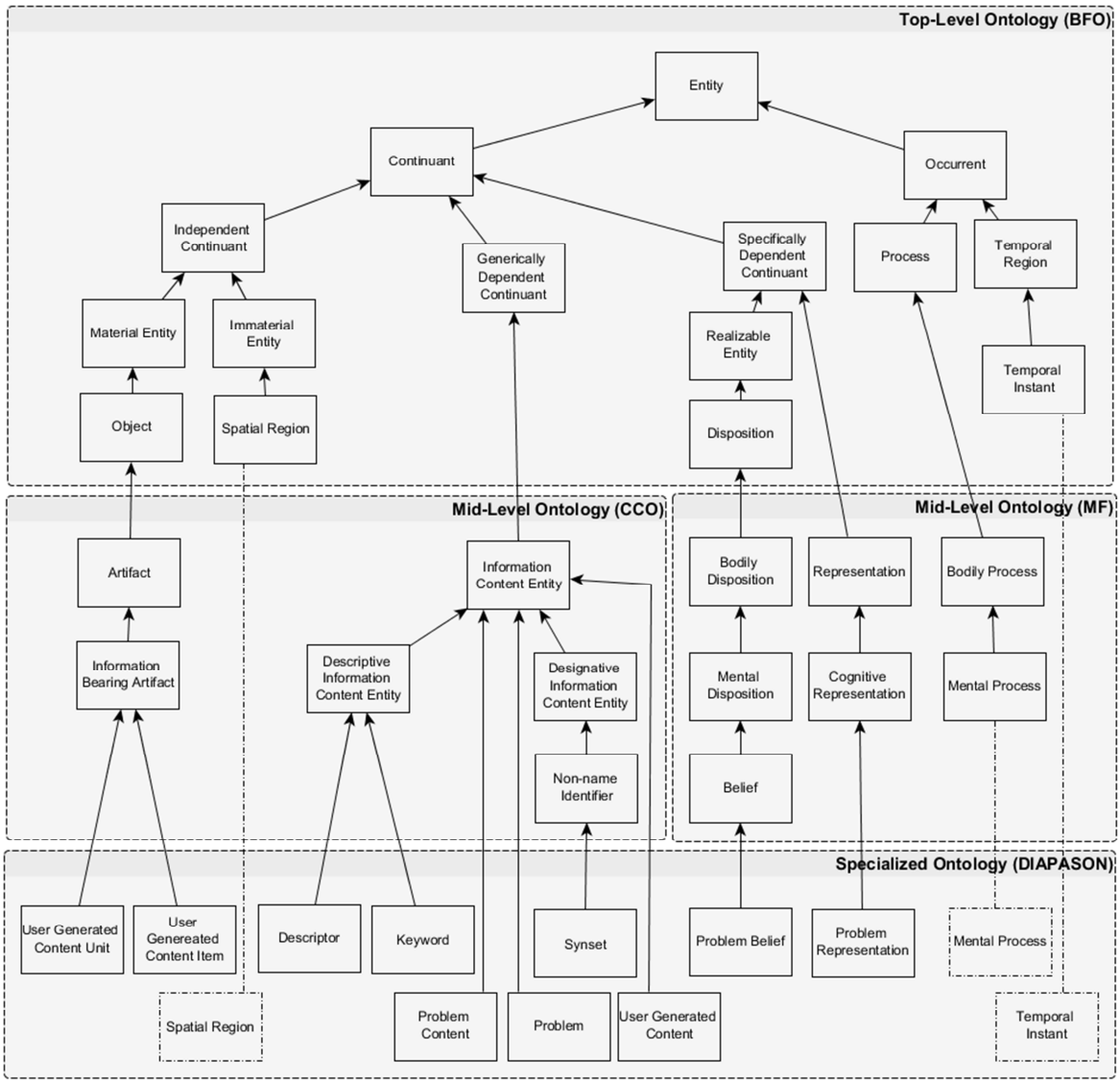

Fig. 6. Aligning the DIAPASON ontology with top- and mid-level ontologies.

carbon dioxide emissions from transport increase health risks. Therefore, dispositional beliefs exist even when we are not actively thinking about them. In contrast, when we are actively thinking about them, we engage in an occurrent belief process during which we take something to be the case.

In the cognitive psychology of memory, activated (occurrent) beliefs are in working memory, whereas latent (dispositional) beliefs are in long-term memory, where both types of beliefs are explicitly represented in the mind (Sperber & Wilson, 2015). Indeed, representationalism is the most prominent theory of belief (McCain, 2016), i.e. a belief involves having a particular representation consciously or stored in one's mind. In this regard, the dual-coding theory (Paivio, 1986) assumes that there are two distinct symbolic subsystems in the human brain, one specialised for non-verbal representation and processing and the other specialised for dealing with language. Therefore, there are two basic modes of representation (i.e. imagistic and verbal), where the nature of their representational

units is different (i.e. *imagens* and *logogens*, respectively). Thus, mental encoding can be done with a sequence of concrete verbal units (e.g. words) in the case of propositional representations.

The philosophical distinction between dispositional and occurrent beliefs has also permeated the design of realism-based ontologies. For example, Barton et al. (2018) and Toyoshima et al. (2020) reflected the dual nature of beliefs in their preliminary description of the foundations for an ontology of belief, desire, and intention. According to these authors, a dispositional belief is a subclass of BFO:DISPOSITION and an occurrent belief is a subclass of BFO:PROCESS, where a given mental process is the trigger of an instance of a dispositional belief that is realised in an instance of an occurrent belief. In contrast, only dispositional beliefs are considered in MF (Hastings et al., 2012), i.e. MF:BELIEF is a subclass of MF:MENTALDISPOSITION. MF is based on the framework laid out in Ceusters and Smith (2010), who distinguished between the level of reality, the level of cognitive representations of this reality (e.g. as embodied in beliefs), and the level of publicly accessible concretizations of such cognitive representations. In this framework, dispositional beliefs are triggered by mental processes (MF:MENTALPROCESS), which create or modify cognitive representations (MF:COGNITIVEREPRESENTATION), whose content is presented, shaped or organised to its bearer. This approach was also adopted in the Medication Adherence Behavior Ontology (Sawesi, 2018).

In DIAPASON, PROBLEMBELIEF represents the dispositional belief of a community problem, so it is a subclass of MF:BELIEF (axioma a4). Thus, PROBLEMBELIEF is (a) a realizable entity, i.e. it is realised by MF:MENTALPROCESS such as judging, thinking, remembering, etc. (axiom a24), and (b) specifically depends on some independent continuant, i.e. a mental functioning related anatomical structure in terms of MF.[5] It should be noted that community problems can only arise when citizens make claims about them, for example, when UGC units are published. In this case, when a UGC is created, the individual consciously holds an occurrent belief. However, if community problems are considered occurrent beliefs, then problems would exist only at the moment they are consciously realised. In other words, as soon as we moved to the next topic, the problem would cease to exist. Therefore, as we assume that ocurrent beliefs are evidence for dispositional beliefs in claimsmaking:

> [...] if a subject has an occurrent belief that *p*, then she thereby has the dispositional belief that *p*. After all, if she is consciously endorsing the content, then she has the information available to mind (she could hardly consciously endorse what was unavailable to her mind) (Rose and Schaffer, 2013, p. 23)

only dispositional beliefs are pertinent to the purpose of DIAPASON, so that cognitive representations are stored in the brain and persist through time past the moment of their acquisition, being ready to be activated when needed. In other words, PROBLEMBELIEF is triggered by at least one MF:MENTALPROCESS (axiom a25), where cco:isTriggerOf and cco:hasTrigger are inverse object properties (axioms a39 and a48). In turn, MF:MENTALPROCESS outputs a PROBLEMREPRESENTATION (axiom a24), where cco:isOutputOf and cco:hasOutput are inverse object properties (axioms a38 and a47). Finally, a PROBLEMREPRESENTATION concretizes PROBLEMCONTENT (axiom a23), since cognitive representations concretize information content entities at a particular time (Smith & Ceusters, 2015), where cco:concretizes and cco:isConcretizedAs are inverse object properties (axioms a37 and a46). However, it is important to note that having an inventory of specific types of community problems related to a given domain does not imply that the knowledge base has an inventory of problem beliefs but of problem themes. Thus, each PROBLEMCONTENT has at least one problem theme, which is

[5] In MF, a mental functioning related anatomical structure (e.g. a particular group of receptors and neurotransmitters in the brain) is that part of an organism that bears a disposition to be the agent of a particular mental process. When a disposition ceases to exist, then its bearer (i.e. receptors and neurotransmitters) will physically change.

categorised as PROBLEM (axiom a22). This decision was influenced by the Theme Ontology (TO) (Sheridan et al., 2019), whose scope is the representation of themes that can be expected to arise in works of fiction (e.g. novels, films, television dramas, comics, etc.). In TO, the concept of "theme" encompasses both the topic treated by a work of fiction (e.g. love in Shakespeare's *Sonnet 116*) and the opinion or judgment conveyed by the work about the topic treated therein (e.g. love remains constant even when assaulted by tempestuous events or by time). In TO, the theme, a subclass of information content entity, is part of a fictional content entity.

On the other hand, CCO comprises twelve ontologies that provide semantics for concepts used in broad domains of interest. Types from two CCO modules were reused: Artifact Ontology and Information Entity Ontology. Our approach to UGC distinguishes between information artefact and information content. CCO:INFORMATIONBEARINGARTIFACT from Artifact Ontology represents an artefact that carries an information content entity and is designed to do so using a particular format or structure. USERGENERATEDCONTENTUNIT and USERGENERATEDCONTENTITEM were created as subclasses of CCO:INFORMATIONBEARINGARTIFACT (axioms a6 and a7), but there can be no CCO:INFORMATIONBEARINGARTIFACT that can be both USERGENERATEDCONTENTUNIT and USERGENERATEDCONTENTITEM (axiom a19). USERGENERATEDCONTENTUNIT, where a numerical identifier is used as a value of the datatype property hasId (axiom a53), can have one or more instances of USERGENERATEDCONTENTITEM as their parts (axiom a27), where TEXTMESSAGE, VOICEMESSAGE, CCO:IMAGE, and CCO:VIDEO are subtypes of USERGENERATEDCONTENTITEM (axioms a8-a11). For example, a given tweet could include a text and a picture. Each USERGENERATEDCONTENTITEM is part of at least one USERGENERATEDCONTENTUNIT (axiom a28). For example, one particular picture can be included in many tweets. As DIAPASON only deals with text processing, we focused on TEXTMESSAGE, which specifies the text and the language as datatype properties (axiom a54). However, we considered all these subtypes of UGC items because ALLEGRO relies on the same ontological model to process them. Indeed, VOICEMESSAGE and CCO:IMAGE are core elements of other ALLEGRO modules, such as SOUND and ADAGIO, respectively. CCO:VIDEO is out of the scope of the current project.

CCO:INFORMATIONCONTENTENTITY from Information Entity Ontology represents a dependent entity that stands in a relation of "aboutness" to some entity, so information content entities cannot exist without some information bearer. USERGENERATEDCONTENT was created as a subclass of CCO:INFORMATIONCONTENTENTITY (axiom a12). Therefore, every USERGENERATEDCONTENTITEM is the carrier of only one USERGENERATEDCONTENT (axiom a28), but every USERGENERATEDCONTENT generically depends on at least one USERGENERATEDCONTENTITEM (axiom a30), where cco:isCarrierOf and cco:genericallyDependsOn are inverse object properties (axioms a40 and a49). For example, whenever identical text is found in multiple retweets, a single instance of USERGENERATEDCONTENT generically depends on multiple instances of USERGENERATEDCONTENTITEM.

Moreover, three other types of CCO:INFORMATIONCONTENTENTITY created for DIAPASON are DESCRIPTOR, KEYWORD and SYNSET, where the first two are subclasses of CCO:DESCRIPTIVEINFORMATIONCONTENTENTITY (axioms a13 and a14), although there can be no CCO:DESCRIPTIVEINFORMATIONCONTENTENTITY that can be both DESCRIPTOR and KEYWORD (axiom a18), and the last is a subclass of CCO:NONNAMEIDENTIFIER (axiom a15). Every USERGENERATEDCONTENT can be described by a DESCRIPTOR (axiom a29), and every DESCRIPTOR describes at least one USERGENERATEDCONTENT (axiom a32), where cco:describes and cco:isDescribedBy are inverse object properties (axioms a35 and a44). For example, when a tweet is not about a problem, there is no descriptor; however, a given descriptor can describe one or more tweets about the same problem type. Descriptors in the DIAPASON ontology were modelled based on WordNet (Fellbaum, 1998), a lexical database where words are organised into sets of synonyms

(synsets); moreover, each synset, which represents a distinct concept, is connected to other synsets through lexical and semantic relations. WordNet can be found as an RDF/OWL resource, i.e. WordNet Ontology,[6] where the main classes are SYNSET, WORDSENSE and WORD and the first two classes have subclasses for the lexical groups present in WordNet, i.e. nouns, verbs, adjectives and adverbs. The classes SYNSET and WORD, together with the datatype properties synsetId for the former (axiom a56) and lexicalForm for the latter (axiom a55), are relevant for the DIAPASON ontology. However, the name KEYWORD was preferred to characterise a representative word of the UGC. Moreover, since DIAPASON can process English and Spanish texts, a datatype property was created to specify the language of the keyword. Therefore, every DESCRIPTOR consists of two parts: one SYNSET and one KEYWORD for every language (i.e. two keywords) (axiom a32), where the actual word/expression and its corresponding language (i.e. "en" for English or "es" for Spanish)[7] are specified using the datatype properties hasLexicalForm and IsLanguage, respectively (axiom a55). When a particular keyword is used in different instances of UGC, it always has the same sense; therefore, every KEYWORD is assigned to only one DESCRIPTOR (axiom a33). Likewise, every SYNSET is assigned to only one DESCRIPTOR (axiom a34), as the synset is a numerical reference (specified as the value of the datatype property hasSynsetId) of the sense of the keywords that are part of the descriptor.

Finally, we relied on BFO to ensure interoperability with other existing and future BFO-compliant ontologies. Although BFO has primarily been developed to provide a top-level ontology for scientific domains, it proved adequate for our aims. In this regard, two BFO classes are used in the content ODP: BFO:TEMPORALINSTANT and BFO:SPATIALREGION. It is worth mentioning that CCO includes Time Ontology, which extends BFO:TEMPORALREGION to describe when events occur, and Geospatial Ontology, which extends BFO:MATERIALENTITY and BFO:SITE to describe the locations of agents and occurrences of events. However, coarse-grained classes such as BFO:TEMPORALINSTANT and BFO:SPATIALREGION are used as they are sufficient for our purposes. In this regard, a particular instance of USERGENERATEDCONTENTUNIT is created on one BFO:TEMPORALINSTANT, where the datatype property hasDateTime holds the specific date and time of its creation (axiom a58), but the same BFO:TEMPORALINSTANT can be the date and time of one or more instances of USERGENERATEDCONTENTUNIT (axiom a26). Therefore, a PROBLEMBELIEF is instantiated on at least one BFO:TEMPORALINSTANT, where each BFO:TEMPORALINSTANT is the temporal region of at least one PROBLEMBELIEF (axiom a26). Moreover, a particular instance of USERGENERATEDCONTENT can mention a named entity as an instance of BFO:SPATIALREGION (axiom a29), where the datatype property hasName keeps its value (axiom a57), but other instances of USERGENERATEDCONTENT can also mention the same named entity (axiom a31), where cco:isMentionOf and cco:isMentionedBy are inverse object properties (axioms a41 and a50).

The model was also extended with generic rules interpreted under standard first-order semantics:

(r1) PROBLEMBELIEF(?p) ∧ BFO:TEMPORALINSTANT(?t) ∧ USERGENERATEDCONTENTUNIT(?u) ∧ isDateTimeOf(?t,?u) → isDateTimeOf(?t,?p)

(r2) PROBLEMCONTENT(?p) ∧ USERGENERATEDCONTENT(?c) ∧ BFO:SPATIALREGION(?s) ∧ cco:isMentionedBy(?s,?c) → cco:isMentionedBy(?s,?p)

(r3) PROBLEMCONTENT(?p) ∧ USERGENERATEDCONTENT(?c) ∧ DESCRIPTOR(?d) ∧ cco:isDescribedBy(?c,?d) → cco:isDescribedBy(?p,?d)

[6] https://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/

[7] The value of the datatype xsd:language is one of the codes defined by RFC 1766, that is, a two-letter language code (taken from the ISO 639 standard) followed optionally by a two-letter country code (taken from the ISO 3166 standard).

Rule (r1) states that if a UGC unit is created on a particular temporal instant, then the problem belief is instantiated on that temporal instant. Rule (r2) describes that if a UGC mentions a given locative reference, then the problem content also mentions the locative reference. Rule (r3) states that if a particular descriptor describes a UGC, then the descriptor also describes the problem content. Again, these rules were constructed because of the subjectivist approach to community problems, where the perception of such problems requires that people convey their beliefs through claimsmaking. Therefore, DIAPASON explores UGC items to reconstruct the content of problem beliefs. For this reason, the instance of PROBLEMCONTENT should be derived from the instance of USERGENERATEDCONTENT, where both share the same named entities and descriptors. It should be noted that these rules can infer true information because the DIAPASON text processor can recognise UGC that expresses factual truth at the time it was contributed, where negative, past tense, and irrealis markers play a significant role in text processing (see Section 4.2).

It should also be recalled that the current purpose of the DIAPASON ontology is the topic categorisation and keyword recognition with short texts as UGC items. To this end, the ODPs in Fig. 4 and Fig. 5 have sufficient breadth and generalization. However, the next stage of the ALLEGRO project, which is out of the scope of this article, is aimed at providing a deeper understanding of citizens' concerns. In particular, we intend to model each of the thirty-four problem domains as a content ODP, which will be constructed from the components of the schemas linked to the problem types of the domain under analysis. A critical mass of problem types for a given domain provides a controlled vocabulary in the form of synsets and facilitates the identification of conceptual patterns across different domain-related problem types. In this regard, the descriptors of problem types can be viewed from a different ontological perspective, i.e. not representing information content entities but the referents of the information content. As a result, for example, the keyword *faded* related to the synset 300404961 in the schema of the problem type ROADMARKING would be converted into the class FADED, which in turn would be modelled as BFO:QUALITY of LINE. This new ontological approach would address each problem type as a new local modelling issue, where current research results would contribute to reducing the complexity of such a task. After defining the problem-type patterns of a given domain, we could construct a module that links such patterns together to provide the domain model.

3.5. Stage A: Implementation

The ontology was programmatically implemented with RDFSharp, an open-source API library for C# to create applications, services and websites capable of modelling, storing and requesting RDF data.[8] RDFSharp allowed implementing a reasoner with forward-chaining materialisation of ontology inferences. In particular, to produce semantic inferences and logical derivations from existing knowledge, the ontology reasoner was modelled using two types of reasoning rules: standard OWL2-DL rules (e.g. entailments based on the inverse property or the transitivity in the SubClassOf taxonomy, among others) and custom SWRL rules, i.e. those corresponding to the rules (r1)-(r3) in Section 3.4.

3.6. Stage B: Specification and Knowledge Acquisition

The second stage aims to construct specific problem types in the different domains. Knowledge sources are similar to those employed at the first stage, but existing data on the Internet (e.g. newspaper headlines and UGC) are also examined. From such knowledge sources, an inventory of problem types related to a given domain is created, and then each specific problem type is described through a

[8] https://github.com/mdesalvo/RDFSharp

statement expressed in English. It should be recalled that, although a single statement is constructed for each problem type, each problem type can be assigned to several domains. Moreover, problem statements should be described in a precise, concise and complete way. On the one hand, the problem statement should be precise, only containing true information for the corresponding problem type. On the other hand, the problem statement should be as concise as possible so that every term is relevant (i.e. specific to the problem being described) and there are no redundant terms (i.e. closely related terms in meaning, such as synonyms and hypernyms, are discarded). Finally, the problem statement should be as complete as possible to cover all the relevant terms for the given problem type.

3.7. Stage B: Conceptualisation and Integration

3.7.1. Conceptualising and integrating the specific problem type

The axiomatisation of the specific problem type involves creating a new class to be integrated into the subClassOf hierarchy of the ontological model. To illustrate, consider SEAWATERQUALITY, BEACHSECURITY and DISAPPOINTMENTMONARCHY, where the first two are problem types related to the quality of urban beaches. Although beach-related problem types primarily pertain to the COMMUNITYFACILITY domain in the INFRASTRUCTURE dimension, some are also linked to other domains. For example, SEAWATERQUALITY can be regarded as a risk factor for the individual's health in case of bathing, so this problem type is also linked to the HEALTH domain in the LIVING dimension. BEACHSECURITY is also linked to the CRIMEJUSTICE domain in LIVING, since law-and-order problems can result from an insufficient number of law-enforcement agents. The corresponding DL axioms are as follows:

$$\text{SEAWATERQUALITY} \sqsubseteq \text{COMMUNITYFACILITY} \sqcap \text{HEALTH}$$
$$\text{BEACHSECURITY} \sqsubseteq \text{COMMUNITYFACILITY} \sqcap \text{CRIMEJUSTICE}$$
$$\text{DISAPPOINTMENTMONARCHY} \sqsubseteq \text{POLITICALSYSTEM}$$

3.7.2. Conceptualising the problem schema

The Specification activity at Stage B returned unstructured knowledge as problem statements. The next step focuses on converting this knowledge into structured conceptual representations (i.e. problem schemas) for problem-type understanding. In this case, *understanding* involves the process of interpretation of UGC, which consists in mapping natural-language text into problem schemas, which will be subsequently used to perform tasks of social analysis. Therefore, we can say that DIAPASON understands a given UGC unit when the system can detect whether or not the UGC unit reflects a problem and recognise the described problem type. In this context, problem understanding focuses on assessing citizens' different perceptions about community problems, where such perceptions can be defined as 'observations noted by individuals, but modulated by their personal feelings, beliefs and perspectives' (Doran et al., 2016, p. 61). Therefore, problem recognition is the task of modelling perceptions embodied in the words and phrases that compose UGC. It can be concluded that problem schemas should contain the shared conceptual knowledge the community has about a given problem type, realised by shared linguistic expressions.

3.7.2.1. Constructions and communicative functions

UGC is expressed in written-down spoken language, so the speech act theory is central to opinion mining. Indeed:

> there seems to be a need to understand *opinions* [...] as expressions of *attitude* or *positive and negative feelings* in a broader sense than *sentiment as polarisation of parts of speech*,

> especially since identification of adequate constructions (grammatical patterns used to express given speech acts) enable coding of such information (Pluwak, 2016, p. 20)

One of the key points in the speech act theory, which is usually attributed to Austin (1962)—and whose ideas were refined by Searle (1969), is that any utterance can have three types of force: locutionary force, i.e. the literal meaning of the words in the utterance, illocutionary force, i.e. the action the speaker is intended to perform through the words, and perlocutionary force, i.e. the actual effect of the words on the hearer. In this respect, the representation of specific problem types focuses on the locutionary and illocutionary forces of text-based UGC.

Some research studies have developed the so-called tweet-act classifiers in the last decade. Such systems intend to automatically recognise speech acts in tweets by not only employing machine learning (Zhang et al., 2011; Vosoughi & Roy, 2016) or deep learning techniques (Algotiml et al., 2019; Saha et al., 2020) but also incorporating handcrafted features into the model, such as lexical features (e.g. distinctive words and phrases), syntactic features (e.g. punctuation marks), and structural features (e.g. POS tags). Researchers primarily follow Searle's typology of speech acts (i.e. *representatives, directives, commissives, expressives,* and *declarations*) to recognise speech acts in tweets. However, the original taxonomy has been progressively modified to make it suitable for speech acts on Twitter. First, Zhang et al. (2011) differentiated between *question* and *suggestion* (i.e. two types of *directives*) but merged *commissives* and *declaratives* into the *miscellaneous* category. Second, Vosoughi & Roy (2016) introduced *request* as a third type of *directives*. Finally, Saha et al. (2020) incorporated *threat* as a single category, keeping the remaining types of *commissives* in *miscellaneous*. All the same, a typology of seven tweet acts is inadequate to categorise citizens' problems such as the fear of crime and the disengagement with the government or political parties, among others, which require a more extensive inventory of negative emotions (e.g. anger, fear, sadness, etc.) and attitudes (e.g. dislike, distrust, pessimism, etc.). A solution to this issue can be found in the way that sophisticated intelligent agents understand language:

> One way is to record in the lexicon a large inventory of canonical patterns for expressing indirect speech acts. For example, among the dozens of frequent paraphrases for asking someone to do X are *I'd like to ask you to do X, Would you mind doing X?, It would be (really) great if you'd do X*, and so on. Since formulas like these are known by people, we must make them known to our agents as well. (McShane, 2017, p. 53)

For this reason, language learning resources were leveraged not only to start getting the knowledge about speech acts in the form of formulaic constructions but also to organise such knowledge based on the notion of *communicative functions* (Wilkins, 1976), which can be regarded as 'categories of pragmatic meaning' in the speech act theory (Peterwagner, 2005, p. 234). For example, Blundell et al. (1982) presented a multitude of phrases for 140 communicative functions, being classified into types such as informational (i.e. seeking and expressing factual information), attitudinal (i.e. asking about and expressing feelings and opinions about something), actional (i.e. committing oneself/others to or asking for a future course of action), and social formulaic (i.e. confirming social relationships), among others. As described in the following subsection, this function-based approach to UGC was adopted to construct the illocutionary part of problem schemas.

3.7.2.2. The notation of problem schemas

Problem schemas consist of two main components: the illocutionary part, which conveys non-propositional meaning, and the locutionary part, which conveys factual information, being both components separated by a colon, as shown in (1).

(1) [illocutionary part] : [locutionary part]

The main building blocks of the illocutionary part of the problem schema are functions, whereas those of the locutionary part are concepts and named entities. To illustrate, the problem schemas corresponding to SEAWATERQUALITY, BEACHSECURITY and DISAPPOINTMENTMONARCHY are shown in (2), (3) and (4), respectively.

(2) ((faeces-114854262 | sewage-114856893) & sea-109426788)
(3) ((P-police-108209687 ^ N-police-108209687) & beach-109217230)
(4) (DISAPPOINTED | PESSIMISTIC) : (monarchy-108363812 | $monarch)

Therefore, problem schemas can assess the intent and the content of online user-generated texts, as shown in (4). It should be noted, however, that the illocutionary component of the problem schema can be irrelevant for the recognition of some problem types; in this case, the descriptive function is ignored, resulting in problem schemas with only propositional content, as shown in (2) and (3). In other words, the locutionary part of the schema is sometimes sufficient to recognise some problems.

Each part of the problem schema contains simple elements (i.e. concepts, functions, named entities, and operators) and complex elements (i.e. expressions), described as follows:

a) Concepts take the form of WordNet synsets, each with an English word. For example, 114854262 represents a synset in (2), whose sense 'solid excretory product evacuated from the bowels' can be linguistically realised by *faeces* and other lexical units in English. It should be noted that English words are expendable in problem schemas as lexical units in many languages can be automatically retrieved from synsets; however, they were kept to facilitate readability so that researchers can manage such knowledge representations more easily.
b) Functions are categories of pragmatic meaning that can be linguistically realised through formulaic constructions, i.e. lexical, syntactic and semantic patterns that can be implemented in various linguistic realisations. For example, DISAPPOINTED and PESSIMISTIC in (4) are linguistically projected to constructions such as *it's a real pity...* and *I'm sceptical about...*, respectively, among many others. For the recognition of community problems, sixteen communicative functions were initially selected from Blundell et al. (1982): one informational function (e.g. saying something is not correct), ten attitudinal functions (e.g. saying you are disappointed, expressing dislikes, or disagreeing), and five actional functions (e.g. warning someone or complaining). However, it was soon realised that there was a need to include not only new functions (e.g. saying you do not have confidence in someone or something) but also new patterns for existing functions since the language used in digital media is characterised by departing from the commonly accepted standard for written text.
c) Named entities are real-world entities (e.g. individuals, locations, and organisations) denoted by proper names. When categorising named entities in problem schemas, categories representing individual instances (e.g. $European_Union for European Union) are differentiated from those representing a set of instances (e.g. $politician for Boris Johnson, Jeremy Corbyn, and Nigel Farage, among many others). For example, $monarch in (4) can be instantiated by the named entity of any member of the royal family of a country, e.g. Elizabeth II, King Charles III, and others in the United Kingdom.
d) There are two types of operators. On the one hand, conceptual operators, e.g. modifiers (M, P) or negation (N), act on the concepts of a proposition to give greater semantic specificity to the description of the problem type. Modifiers aim to present a particular quality, entity, event or

situation at a higher (M) or lower (P) quantity, degree or intensity than expected for the state of affairs being described, where the referents modified can be nouns (e.g. *many/few* people), verbs (e.g. eat a *lot/little*), or adjectives (e.g. *highly/slightly* toxic). In contrast, negation (N) refers to any construct that introduces the lack of a given quality, entity, event or situation. Like functions, conceptual operators are projected to linguistic realisations; for example, the modifier P and the negation N in (3) can be linguistically mapped to *a few of...* and *there aren't...*, respectively, just to mention a few. On the other hand, logical operators, i.e. conjunction (&), inclusive disjunction (|) and exclusive disjunction (^), connect two or more elements of the same kind.

e) Expressions hold a set of concepts, functions, named entities or other expressions, providing that such elements are concatenated with logical operators. Expressions employ round brackets to determine their scope.

3.7.2.3. The role of problem schemas

Problem schemas play a critical role when DIAPASON performs topic categorisation and keyword recognition. On the one hand, keywords are selected from the words linked to the synsets in problem schemas. According to Beliga et al. (2015), keyword-selection methods can be divided into two categories: keywords can be selected from a predefined controlled vocabulary of terms (i.e. keyword assignment) or directly from the source document (i.e. keyword extraction). In DIAPASON, keyword recognition adopts the former approach, as problem schemas provide the system with a collection of synsets from which keywords can be derived based on the relevance to the text content. In other words, the synsets in problem schemas that are semantically related to lexical units in the UGC are instantiated as keywords. Therefore, this approach can assign keywords that are not present in the text, which Zhang & Xu (2009) called "implicit keywords". From an ontological perspective, the problem schema of a problem type provides a finite set of literals that constrain the values of the datatype properties hasSynsetId and hasLexicalForm in the classes SYNSET and KEYWORD, respectively. Section 4.2.5 illustrates how keyword recognition is performed with problem schemas.

On the other hand, topic categorisation also relies on problem schemas. In particular, they contribute to compiling a robust training dataset, which DIAPASON employs to classify UGC through a supervised deep-learning model. To this end, problem schemas have a twofold purpose. They can serve as a provider of domain terms used to explore the Web (e.g. seed terms for the Twitter search API) and thus construct a UGC corpus, where instances can then be manually annotated as part of the training dataset. In addition, problem schemas can be leveraged to create a synthetic training dataset or expand an existing training dataset with synthetic data, i.e. pseudo-realistic data samples that capture the diversity of naturally constructed occurrences, so that the performance of the system can be improved. Indeed, we intend to build a training dataset consisting of both real-world and artificial instances of linguistic data. This hybrid approach aims to align the notion of E-language (external language) with that of I-Language (internal language), a distinction introduced by Chomsky (1986). On the one hand, the representative sample of written utterances in a corpus represents an "external" object. However, this raises the problem that this object cannot be constituted by the totality of utterances made in a speech community, not to mention that a combination of words that has never been used could be uttered someday. On the other hand, the linguistic knowledge (e.g. a network of semantic associations) that resides in speakers' brains and underlies their linguistic performance can be considered an "internal" object. It is precisely this knowledge that allows people not only to produce utterances that others can understand but also to understand the utterances that others have made. Taylor (2012) supported the idea that both models are closely aligned, thus proposing a relation between language as it is encountered in the external world and language as it is mentally represented. This is the relation on

which hybrid training datasets can be grounded. The remainder of this section provides a detailed account of how synthetic training datasets can be constructed from problem schemas.

Using problem schemas as a data generator for the training dataset involves creating two main types of constructs: conceptual profiles and conceptual contexts. First, conceptual profiles were constructed as follows:

1- We obtained a list of all the WordNet synsets included in problem schemas, which we call "source synsets".

2- Each source synset was expanded through different WordNet relations based on the category of the source synset: *hyponym, pertain_to,* and *related_to* for nouns, *hyponym, subevent,* and *related_to* for verbs, *near_synonym, pertain_to, is_derived_from,* and *related_to* for adjectives, and *is_derived_from* for adverbs. We call "expanded synsets" to such semantically related synsets detected from source synsets.

3- We obtained synset paraphrases for each source synset and its expanded synsets, where the synsets discovered in this step were also considered as expanded synsets. Synset paraphrases were retrieved from our Synset-based Paraphrase Database (SPD), which was automatically constructed by leveraging resources such as the Paraphrase Database (Ganitkevitch et al., 2013), LessLex (Colla et al., 2020), and WordNet.

As a result, the conceptual profile generated from a given source synset takes the form of an array of expanded synsets together with the source synset. To illustrate, suppose we take SEAWATERQUALITY, whose problem schema was presented in (2). The conceptual profile of each source synset in this problem schema is presented in (5).

(5) 114854262 [faeces-114854262, dog_shit-109268480, droppings-114854847, faecal-303065685, defecate-200074038, etc.]
114856893 [sewage-114856893, 104179126-sewer, drainage-100396029]
109426788 [sea-109426788, 109376198-ocean, 300463399-marine, 302887899-marine, 300124353-marine, etc.]

It should be noted that only synsets are part of conceptual profiles; however, we include lexical units in these examples to facilitate readability. Indeed, the conceptual nature of this type of constructs allows different senses of a given word to be related within the same profile, as illustrated in 109426788, which includes various senses of marine, such as 'bordering on or living or characteristic of those near the sea' [300463399], 'of or relating to the sea' [302887899], and 'native to or inhabiting the sea' [300124353].

Second, conceptual contexts were constructed for each problem schema as follows:

1- As the problem schema is a complex logical expression, a truth table is generated to obtain all true logical statements, with each statement considered a conceptual context. For example, the statements in (6) were derived from the problem schema in (2).

(6) faeces-114854262 & sea-109426788
sewage-114856893 & sea-109426788
faeces-114854262 & sewage-114856893 & sea-109426788

2- Each conceptual context directly derived from the problem schema (i.e. source conceptual context) can generate other conceptual contexts in a data-augmentation process (i.e. expanded conceptual context). Following Wei & Zou (2019), automatic data augmentation was performed through three techniques that were sequentially applied: synset replacement, i.e. any synset in the conceptual context is replaced with one of its conceptual paraphrases, synset swap, i.e. two randomly chosen synsets in the sentence swap their positions, and synset insertion, i.e. a random paraphrase of a random synset in the conceptual context is inserted into a random position in the context. Different resources were leveraged in these techniques. For example, synsets in conceptual contexts were replaced with synsets from conceptual profiles for synset replacement, and syntagmatically related synsets were taken from SyntagNet (Maru et al., 2019) for synset insertion.

Following our example, some of the expanded conceptual contexts derived from (6) are presented in (7).

(7) sewage-114856893 & marine-300463399
faecal-303065685 & sewer-104179126 & sea-109426788
marine-300124353 & faeces-114854262
drainage-100396029 & sea-109426788 & coast-109428293
sea-109426788 & defecate-200074038

As explained in Section 4, conceptual contexts play a pivotal role not only in constructing the synthetic training dataset but also in the form in which the test dataset is fed into the deep-learning model.

3.8. Stage B: Implementation

DIAPASON includes an ontology-development environment to help researchers conceptualise problem types in an assisted fashion. In particular, this environment provides several utilities, such as an editor to add, remove or modify problem types, a browser to inspect the information stored in the ontology, and a searcher to look for specific information, among others. In the case of problem schemas, a validator checks the syntax of representations.

Therefore, a new specific problem type can be created and integrated into the subClassOf hierarchy, and then the problem schema resulting from the Conceptualisation activity is entered into the database. Subsequently, the new corresponding DL axioms are automatically incorporated into the existing RDF model.

3.9. Stage B: Evaluation

The DIAPASON ontology is populated with individuals and object property assertions about these individuals (A-Box) derived from processing a UGC corpus with problem schemas. The instantiation of the ontology provides an opportunity for intrinsic and extrinsic evaluation. On the one hand, the ontology was validated against a predefined set of rules analysing the model to detect semantic inconsistencies and violations that could lead to the materialisation of incorrect knowledge. In particular, RDFSharp is provided with sixteen built-in rules to verify that, for example, assertions must respect domain/range constraints or that restrictions must be fully defined, among other types of validation.

On the other hand, the ontology was transformed into an RDF graph with RDFSharp to execute high-level SPARQL queries. Therefore, the usefulness of the ontology was tested by running SPARQL

queries based on the competency questions. For example, the SPARQL query constructed to answer the competency question (ii) is presented in (8).

(8) SELECT DISTINCT ?w WHERE {?p :isCarrierOf ?q . ?q :isDescribedBy ?x . ?x describes ?y . ?y hasPart ?z . ?z rdf:type :EcologicalHazard . ?p :hasText ?w}

It should be noted that the content of the RDF graph can be serialised into an RDF/XML document for further exploration of the DIAPASON ontology with external tools.

Stage B is in continuous development, as new types of specific problems can arise anytime anywhere, resulting in an ongoing review of knowledge in the problem-type level of the DIAPASON ontology. Indeed, the permanent updating of ontologies is an intrinsic property of such resources since any ontology is an information object representing the time, place and cultural environment in which it was created (Brewster et al., 2004).

4. Topic categorisation and keyword recognition with DIAPASON

4.1. Materials

We experimented with two small test datasets to illustrate the role of the ontology in problem detection, which involves the NLP tasks of topic categorisation and keyword recognition, and the subsequent construction of the knowledge base. For this experiment, we explored text-based UGC items for urban sensing, where local government websites and smartphone applications allow citizens to report non-emergency problems about urban infrastructure and services, as such issues can decrease citizens' QoL within their communities. In particular, the test datasets were constructed from citizens' requests stored in the Open311 system in the City of Bloomington (USA):[9]

a) Test Dataset A (667 messages), which consists of requests that fall under the original categories of "Traffic-Related Complaints", "Traffic Signals", and "Traffic Suggestions", and
b) Test Dataset B (1391 messages), which includes the messages in Test Dataset A apart from requests that pertain to the original categories of "Animal Control" and "Sanitary Sewers".

We performed multi-class text classification with both test datasets. However, for experimental purposes, we deliberately decided that the number of classes and their thematic granularity should be different for each dataset. In Test Dataset A, three annotators with linguistic background were invited to manually label the messages with the fine-grained classes corresponding to the following problem types: PARKING, ROADFURNITURE, ROADMARKING, ROADSAFETY, ROADSURFACE, STREETFURNITURE, and TRAFFICCONGESTION. In contrast, Test Dataset B was labelled with three coarse-grained classes, i.e. Animal, Sewer, and Traffic, where the first two correspond to the problem types ANIMALCONTROL and SEWER. Moreover, the same three annotators manually extracted the keywords for the UGC items in both test datasets. The annotators were instructed on how to detect candidate keywords. In particular, they identified nouns, verbs and adjectives found in the text of each UGC item that contribute to constructing "the abridged account of the problem" described by the message. Then, each lexical candidate was tagged with a synset from the problem schema of the problem type corresponding to the UGC item. The annotators attempted to solve the cases of

[9] http://www.open311.org

disagreement by discussion so that the test datasets consisted of messages labelled with classes, keywords and synsets recommended by at least two annotators.

4.2. Methodology

4.2.1. Overview

DIAPASON employs a text-processing module to derive extensional knowledge from a UGC corpus uploaded through the online workbench and thus construct the knowledge base. Figure 7 shows the architecture of this module, which takes the form of a pipeline of NLP and text-mining tasks structured into four stages (i.e. text pre-processing, text processing, text classification, and keyword recognition), where the final outcome serves to populate the ontology.

4.2.2. Pre-processing texts

This stage involves the spelling and typographical standardisation of texts. For example, in the case of tweets, emojis are converted into lexical units, hashtags are segmented into tokens, and references and URL links are automatically removed, among other standardisation methods. In this experiment, this stage primarily focused on capitalization and spelling-error correction.

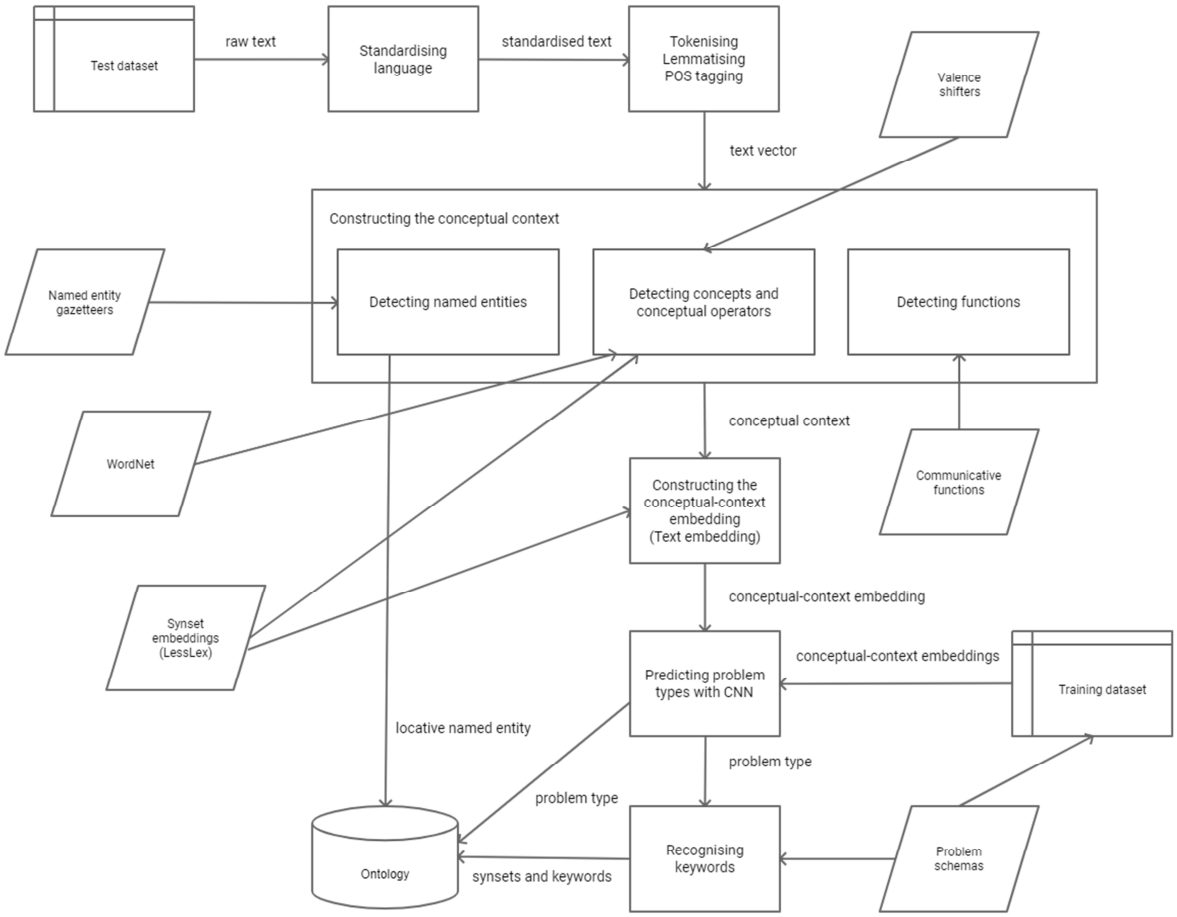

Fig. 7. DIAPASON text processor.

4.2.3. Processing texts

Each pre-processed text was split into sentences, each sentence was tokenised, and tokens were POS-tagged and lemmatised. At this point, a UGC item is represented as the vector $T_m = (w_{m1}, w_{m2}, \ldots, w_{mq})$, where w_{mn} represents the object for the n^{th} token that occurs in the text m and q is the total number of objects. In turn, each w_{mn} is defined with attributes such as the position in the text, the word form, the lexeme, and the POS.

The next task consisted in constructing the conceptual context that describes the UGC from this array of objects. Three parallel tasks were performed to achieve this goal, each aimed at detecting the major elements in the conceptual context, i.e. named entities, concepts, and functions:

a) Named entities were detected with lookup-based methods based on gazetteers, thus recognising the names of people, locations, and organisations relevant for each dimension under analysis in the DIAPASON ontology, e.g. names of politicians and political institutions in GOVERNANCE. In the case of locative named entities, we employed nLORE (neural LOcative Reference Extractor) (Fernández-Martínez & Periñán-Pascual, 2021b), a bi-directional Recurrent Neural Network with Long Short Term Memory as a hidden layer structure and a Conditional Random Fields layer on top exploiting the linguistic knowledge provided by LORE (Fernández-Martínez & Periñán-Pascual, 2021a). We can perform fine-grained locative-reference extraction with nLORE, where references range from geopolitical entities (e.g. towns, cities, states, and countries) and natural landforms (e.g. lakes, rivers, mountains, ridges, and beaches) to points of interest (e.g. schools, churches, malls, museums, and police stations) and traffic ways (e.g. streets, avenues, turnpikes, boulevards, highways, and roads). Moreover, nLORE can capture complex locative references, consisting of location-indicative words and locative markers accompanying a given place name, e.g. *35miles from New York, South of Madrid, Dyckman Street Station,* etc.

b) Concepts were detected from the list of unigrams and bigrams generated from the text vector T_m. In the case of unigrams, only nouns, verbs and adjectives were retained, where common words that are likely to be non-informative were discarded. In the case of bigrams, we applied a set of basic POS-based lexical filters; in particular, some classical patterns for bigrams in English are *v + p, n + n,* and *a + n*, where *n, v, a,* and *p* represent nouns, verbs, adjectives, and particles, respectively. The candidate ngrams that had no sense embedding in LessLex corresponding to a WordNet synset were filtered out. We subsequently identified the contextual valence shifters close to the selected ngrams. Valence shifters, which are linguistic structures that can affect the polarity (i.e. negative or positive attitude) expressed by some lexical units, are usually taken into consideration to improve the accuracy of classification in opinion mining and sentiment analysis. In DIAPASON, valence shifters play a key role in determining conceptual operators or filtering out irrelevant concepts. Based on Polanyi & Zaenen (2004), we considered three types of valence shifters: negation cues, modifiers (i.e. intensifiers and diminishers), and irrealis markers. On the one hand, negation cues and modifiers were directly projected as conceptual operators. On the other hand, it should be recalled that we are primarily concerned with present facts, where we define *present* in an inclusive way, i.e. 'if it has existence at the present moment, allowing for the possibility that its existence may also stretch into the past and into the future' (Quirk et al., 1985, p. 175). In the framework of epistemic modalities, Givón (1993) explained that presuppositions and realis assertions express facts, and irrealis and negative assertions express non-facts. Therefore, as negative cues were already projected as conceptual operators, non-present facts were further detected using (a) past-tense markers, i.e. grammatical devices (e.g. simple past and past perfective tenses) that express a state of affairs completed before the present moment, and (b) irrealis markers, such as English

modals (e.g. Unemployment *could* increase to 10%) and conditional clauses (e.g. *If* unemployment increases to 10%...), among others. In such cases, ngrams within the scope of past-tense and irrealis markers were discarded.

c) Functions were detected by adopting a rule-based approach grounded on the lexical, syntactic and semantic features of the input. To this end, we are constructing a repository of generic domain-independent patterns for each communicative function through which linguistic realisations can be generated or parsed.

A conceptual context similar to those found in the training dataset (see Section 3.7.2.3) was constructed from the main elements discovered in the previous tasks. The projection of such elements into conceptual operators, functions, and named entities was straightforward. In contrast, valid ngrams required to be disambiguated so that they could be projected into WordNet synsets. To this end, we performed unsupervised word sense disambiguation (WSD) based on synset embeddings. As in Basile et al. (2014), we employed a variation of the Lesk algorithm where the notion of lexical overlapping was replaced with similarity in a semantic vector space, being semantic relatedness determined by a proximity measure (e.g. cosine similarity). In particular, we applied the LessLex semantic model to WSD for computing the overlap between the meaning of an ngram and its context, both represented as embeddings. Therefore, given a sequence of ngrams (i.e. g_1, g_2, ..., g_q) generated from T_m, we disambiguated each g_i at a time by considering the similarity between the embedding associated with each meaning of g_i (i.e. WordNet synset) and the context embedding, where the context was represented by all the ngrams co-occurring with g_i in the text. The context embedding was constructed by adding the embeddings of all the senses of the ngrams in the context. The meaning with the highest similarity was selected. In this second part of the task, the POS of ngrams helped discard some irrelevant senses.

At the end of this stage, for example, the traffic-related complaint in (9) took the form of the conceptual context in (10).

(9) The crosswalk stripes on the west side of Rogers in the new crosswalk between First and Second streets are gone, perhaps scraped off by a snow plow. Only half the street looks like a crosswalk.

(10) crosswalk-103137228 & stripe-104683136 & street-104334599 & scrape-201308160 & snow-115043763 & plough-202096853

4.2.4. Classifying texts

Identifying which texts describe which problems is regarded as a text-classification task. To this end, we elaborated a training dataset and constructed a two-dimensional convolutional neural network (CNN) model to make predictions on the new data, i.e. classify UGC items in Test Dataset A and Test Dataset B. Each new data sample was fed into the CNN model in the form of a conceptual-context embedding, which was constructed from the embeddings of the elements in the conceptual context returned by the previous stage. In particular, pre-defined 300-dimension embeddings linked to WordNet synsets were directly obtained from LessLex in the case of named entities, concepts, and functions:

a) When a named entity was detected in the conceptual context (e.g. $politician), the system retrieved the synset embedding of the named entity type (e.g. the embedding of the synset 110450303, which corresponds to *politician*).

b) When a concept was found (e.g. street-104334599), the system retrieved the corresponding synset embedding. In the case of conceptual operators, however, we had to search for an effective method for introducing external information into the embedding space. The solution

was found in Madukwe et al. (2021), who explored three methods (i.e. counter-fitting, concatenation, and ensembling) to incorporate emotion in pre-trained word embeddings and thus improve the performance of hate speech detection. In this regard, we chose the concatenation method, which does not require additional training or refining of the embedding space, resulting in a computationally efficient method. In particular, we generated a binary three-dimensional embedding that could reflect the conceptual operator involved with each concept. In this operator embedding, each dimension referred to one of the three conceptual operators (i.e. M, P, and N), where the value 1 indicated the presence of the operator and 0 represented its absence. For example, (11) and (12) are conceptual contexts derived from the problem schema in (3).

(11) (P-police-108209687 & beach-109217230)
(12) (N-police-108209687 & beach-109217230)

The operator embeddings for the concept 108209687 in (11) and (12) are 010 and 001, respectively, whereas the operator embedding for the concept 109217230 is 000.

c) When a function was detected (e.g. DISAPPOINTED), the system recovered the synset embedding of the most representative word (e.g. the embedding of the synset 302333976, which corresponds to *disappointed*).

The conceptual-context embedding was constructed as follows. Suppose that $x_j \in \mathbb{R}_k$ is the k-dimensional embedding corresponding to the j^{th} core element (i.e. named entity, concept or function) in the conceptual context automatically generated from a given UGC. In this experiment, $k = 303$, i.e. the operator embedding was concatenated with the synset embedding of named entities, concepts, and functions; in the case of named entities and functions, the operator embedding was always 000. Following Kim (2014), the embedding of the conceptual context that represents the text m is constructed as $x_m = x_{m,1} \oplus x_{m,2} \oplus ... \oplus x_{m,r}$, where $\oplus$ is the concatenation operator and $r = 6$ in this experiment (i.e. six core elements in the conceptual context). Therefore, any conceptual-context embedding is a vector of 1818 dimensions (i.e. 6 embeddings x 303 dimensions). When the total number of elements occurring in the conceptual context was higher than r, the system took the embeddings of the first six elements. In contrast, when the number of elements was lower than r, we employed zero padding in the resulting conceptual-context embedding to maintain the output size. In any case, the order of occurrence of the elements in the conceptual context is relevant, as CNN requires that the vector representation of data should preserve internal locations in the input.

In this experiment, we created a synthetic training dataset from the knowledge stored in problem schemas, as constructing a real-world annotated corpus for DIAPASON is still a work in progress. In particular, 1000 conceptual contexts were automatically generated for the problem schema of each target class in the test datasets, as described in Section 3.7.2.3. Finally, an embedding was constructed for each conceptual context in the training dataset, following the same procedure as described above.

We employed a simple two-dimensional CNN architecture, whose stack of layers is shown in Fig. 8.

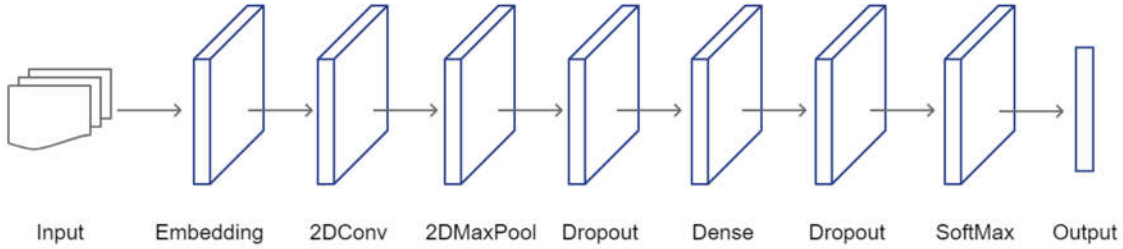

Fig. 8. CNN model for text classification with the DIAPASON ontology.

First, the input data (i.e. test dataset) was represented as a *t* x *d* matrix in the embedding layer, where *t* represents the number of texts (i.e. conceptual contexts) and *d* is the number of dimensions of each text embedding (i.e. conceptual-context embeddings), being *d* = 1818; the depth of this embedding layer was 1. Second, the two-dimension convolutional layer consisted of 32 filters, whose width and height were 5, where the stride size was 1 and the zero-padding for width and height was also 1; the activation function was Rectified Linear Unit (ReLU). Third, the two-dimension MaxPool layer reduced the information by extracting the relevant features that helped make the classification decision; the pool width and height were set to 2. Fourth, we set the dropout rate to 0.5 in each dropout layer, used for regularization to prevent overfitting. Fifth, the fully-connected (or dense) layer involved 1818 neurons, where the activation function was ReLU. Finally, a SoftMax layer of 7 nodes for Test Dataset A and 3 for Test Dataset B, i.e. one for each class, was used for prediction. Regarding the configuration of the general hyperparameters for training our CNN model, we employed Root Mean Square Propagation (RMSProp) as the optimiser method for 10 epochs with a batch size of 256, a learning rate of 0.001, and a momentum coefficient of 0.9.

4.2.5. Recognising keywords

Once text classification has been performed, each concept and named-entity embedding in the conceptual context representing the UGC is compared with each synset vector in the problem schema based on cosine similarity, assuming that the closer an individual concept or named-entity embedding is to a given synset vector, the more likely the synset is to become part of the descriptor. Finally, English and Spanish keywords were obtained from the words linked to the selected synsets.

For example, the conceptual context in (10) was classified as ROADMARKING, whose problem schema is presented in (13).

(13) ((((marking-104680285 ^ line-106799897) & faded-300404961) ^ paint-201362736) & (road-104096066 ^ street-104334599))

Considering the cosine-similarity coefficients computed between each synset in (10) and each synset in (13), we could obtain the highest score for each synset in the problem schema, shown within square brackets in (14).

(14) ((((marking-104680285[0.434] ^ line-106799897[0.732]) & faded-300404961[0.258]) ^ paint-201362736[0.343]) & (road-104096066[0.845] ^ street-104334599[1.0]))

Finally, logical operators and similarity scores helped determine the most salient synsets in the problem schema, as shown in (15).

(15) line-106799897 & faded-300404961 & street-104334599

4.2.6. Populating the ontology

Following the above example, DIAPASON obtained the information in Table 1 from the metadata (TicketID and EnteredDate) and core data (Description) of the UGC unit in the Open311 database. The DL axioms generated from Table 1 are as follows, which serve to construct the A-Box component of the knowledge base:[10]

(a59) ROADMARKING(P1)

(a60) BFO:TEMPORALINSTANT(I1)
(a61) hasDateTime(I1, '2015-04-23T14:07:06')

(a62) BFO:SPATIALREGION(L1)
(a63) hasName (L1, 'First and Second streets')

(a64) DESCRIPTOR(D1)
(a65) SYNSET(S106799897)
(a66) hasSynsetId(S106799897, '106799897')
(a67) KEYWORD(LINE)
(a68) hasLexicalForm(LINE, 'line')
(a69) isLanguage(LINE, 'en')
(a70) KEYWORD(LINEA)
(a71) hasLexicalForm(LINEA, 'línea')
(a72) isLanguage(LINEA, 'es')
(a73) hasSynset(D1, S106799897)
(a74) hasKeyword(D1, LINE)
(a75) hasKeyword(D1, LINEA)

(a76) DESCRIPTOR(D2)
(a77) SYNSET(S300404961)
(a78) hasSynsetId(S300404961, '300404961')
(a79) KEYWORD(FADED)
(a80) hasLexicalForm(FADED, 'faded')

Table 1. UGC information extraction: an example.

ID	144763
Timestamp	2015-04-23T14:07:06
Language	en
Problem type	ROADMARKING
Locative named entity	First and Second streets
Synsets [keywords]	106799897 [line (en), línea (es), ligne (fr), linea (it), etc.] 300404961 [faded (en), descolorido (es), décoloré (fr), scolorito (it), etc.] 104334599 [street (en), calle (es), rue (fr), strada (it), etc.]

[10] Keywords in languages other than English and Spanish have been ignored in the DL axioms.

(a81) isLanguage(FADED, 'en')
(a82) KEYWORD(DESCOLORIDO)
(a83) hasLexicalForm(DESCOLORIDO, 'descolorido')
(a84) isLanguage(DESCOLORIDO, 'es')
(a85) hasSynset(D2, S300404961)
(a86) hasKeyword(D2, FADED)
(a87) hasKeyword(D2, DESCOLORIDO)

(a88) DESCRIPTOR(D3)
(a89) SYNSET(S104334599)
(a90) hasSynsetId(S104334599, '104334599')
(a91) KEYWORD(STREET)
(a92) hasLexicalForm(STREET, 'street')
(a93) isLanguage(STREET, 'en')
(a94) KEYWORD(CALLE)
(a95) hasLexicalForm(CALLE, 'calle')
(a96) isLanguage(CALLE, 'es')
(a97) hasSynset(D3, S104334599)
(a98) hasKeyword(D3, STREET)
(a99) hasKeyword(D3, CALLE)

(a100) USERGENERATEDCONTENTUNIT(U1)
(a101) hasId(U1, '144763')
(a102) isDateTimeOf(I1, U1)
(a103) TEXTMESSAGE(T1)
(a104) hasText(T1, 'The crosswalk stripes on the west side of Rogers in the new crosswalk between First and Second streets are gone, perhaps scraped off by a snow plow. Only half the street looks like a crosswalk.')
(a105) hasLanguage(T1, 'eng')
(a106) bfo:hasPart(U1, T1)
(a107) USERGENERATEDCONTENT(C1)
(a108) cco:isCarrierOf(T1, C1)
(a109) cco:isMentionOf(C1, L1)
(a110) cco:isDescribedBy(C1, D1)
(a111) cco:isDescribedBy(C1, D2)
(a112) cco:isDescribedBy(C1, D3)

(a113) PROBLEMCONTENT(W1)
(a114) cco:isMentionOf(W1, L1)
(a115) cco:isDescribedBy (W1, D1)
(a116) cco:isDescribedBy (W1, D2)
(a117) cco:isDescribedBy (W1, D3)
(a118) bfo:hasPart(W1, P1)
(a119) PROBLEMREPRESENTATION(X1)
(a120) cco:isConcretizedAs(W1, X1)
(a121) MENTALPROCESS(Y1)
(a122) cco:isOutputOf(X1, Y1)
(a123) PROBLEMBELIEF(Z1)

(a124) isTriggerOf(Y1, Z1)
(a125) isDateTimeOf(I1, Z1)

4.3. Results and discussion

On the one hand, DIAPASON provided each text with one of the classes corresponding to the problem types in the test datasets. In this context, good-fit predictive models should perform well both with the data used to train them and with the new data on which predictions will be made. In our case, the challenge is that the model should perfectly learn the synthetic training dataset generated from problem schemas and perform well on the test datasets consisting of real-world UGC texts. On the one hand, we evaluated the CNN model with the training dataset, using 70% of the data for training and 30% for validation. As the prediction error was 0.001, we conclude that the model achieved high accuracy on the training dataset. On the other hand, we evaluated the CNN model with the test datasets. The classification of the messages in Test Dataset A and Test Dataset B provided an accuracy of 0.519 and 0.893, respectively. As expected, Test Dataset B achieved a better result than Test Dataset A, where each class showed a similar word distribution as all citizens' requests deal with traffic issues. It should be noted, however, that both results are positive and promising as there is considerable room for improvement. First, we only employed synthetically generated data (i.e. training dataset) to represent the patterns underlying data extracted from the real world (i.e. test dataset). Natural data augmentation can help the training dataset have the same distribution as the test dataset. However, we intend to continue with a hybrid training dataset, consisting of human-generated messages from UGC items together with synthetically generated data, so the latter can still play a significant role. As the high performance of deep-learning models depends on the quantity and quality of training data, preparing a huge manually crafted annotated dataset could become counterproductive since it is a time-consuming, error-prone and expensive task. Moreover, synthetically generated data can mitigate the imbalanced learning problem in training datasets, where some classes can be underrepresented in comparison to others. Second, we did not perform any dataset-specific tuning of the CNN model that could lead to optimal results; instead, we used standard values for the hyperparameter configuration. Finally, after functional and common words were discarded from the test dataset, we simply relied on the first six elements (i.e. named entity, concept or function) in each new conceptual context to construct the text embedding. Therefore, better performance is likely to be obtained with natural data augmentation, hyperparameter optimisation, and adaptive feature selection.

On the other hand, DIAPASON provided each text with keywords linked to the most salient synsets in the problem schema corresponding to the problem detected. Considering the successfully classified UGC items, synset-based keyword recognition with Test Dataset A and Test Dataset B provided an accuracy of 0.907 and 0.912, respectively.

5. Conclusions

This interdisciplinary study is part of the research project ALLEGRO, where UGC can be analysed to understand the problems that affect citizens in a given community. The project focuses not only on the quality of urban infrastructure and services (e.g. housing conditions, water supply, and traffic, among others) but also on understanding the city's sociological character, which is reflected through people's concerns (e.g. cultural conflicts, unemployment, and violence, to name a few). Therefore, QoL is about not only surrounding conditions but also personal experiences. To this end, social sensing can process, aggregate and organise UGC to produce valuable information that can improve the understanding of psycho-social dynamics. Indeed, smart-city applications can benefit from social sensors, since UGC can

convey eyewitness accounts of social occurrences or reveal the citizens' perceptions of the events and states that can disrupt their QoL.

In this context, DIAPASON, an ALLEGRO module aimed at processing the text message in UGC units, intends to be presented as an online workbench for users (e.g. researchers) to explore and experiment with UGC corpora, where the research goal is the variety of issues that can undermine QoL. The research described in this article demonstrates that DIAPASON can successfully perform topic categorisation and keyword recognition with UGC, particularly with text messages written in English or Spanish. DIAPASON relies on the ontology and the problem schemas of problem types to derive extensional knowledge from such corpora. On the one hand, the DIAPASON ontology provides the semantic data model of the knowledge required to understand problems affecting citizens in their community, which are organised into problem realms, dimensions, and domains. The ontology design was determined by the objective of DIAPASON, i.e. topic categorisation and keyword recognition, so the analysis of UGC results in detecting a PROBLEM that can be summarised by a DESCRIPTOR, which takes the form of SYNSET and KEYWORD. On the other hand, problem schemas are constructed as conceptual representations that play an active role in problem-type detection, i.e. revealing the problem types described in the UGC units. Moreover, problem schemas are considered a critical component in ontology development, as they restrict the synsets and keywords used as descriptors for a given problem type. Finally, the experiment with DIAPASON also demonstrated that NLP, deep learning, and knowledge engineering techniques are integrated from a hybrid approach, which combines symbolic representations of knowledge (e.g. problem schemas) with sub-symbolic models (e.g. CNN), making headway in real-world natural language understanding (i.e. problem-type detection) applications. Therefore, this study provides valuable insight into modelling knowledge resources that contribute to developing smart societies based on the social-sensing paradigm.

Acknowledgements

This article was supported under grant PID2020-112827GB-I00 funded by MCIN/AEI/10.13039/501100011033, and under grant number 101017861 [project SMARTLAGOON] by the European Union's Horizon 2020 research and innovation program.

References

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis, 16*, 3-9.

Algotiml, B., Elmadany, A., & Magdy, W. (2019). Arabic Tweet-Act: Speech act recognition for Arabic asynchronous conversations. *Proceedings of the Fourth Arabic Natural Language Processing Workshop* (pp. 183-191). Association for Computational Linguistics.

An, J., & Weber, I. (2015). Whom should we sense in "social sensing" - analysing which users work best for social media now-casting. *EPJ Data Science, 4*(22), 1-22.

Appio, F. P., Lima, M., & Paroutis, S. (2019). Understanding Smart Cities: Innovation ecosystems, technological advancements, and societal challenges. *Technological Forecasting & Social Change, 142*, 1-14.

Arthur, R., Boulton, C. A., Shotton, H., & Williams, H. T. P. (2018). Social sensing of floods in the UK. *PLoS ONE 13*(1). https://doi.org/10.1371/journal.pone.0189327

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Balaji, B., Bhattacharya, A., Fierro, G., Gao, J., Gluck, J., Hong, D., Johansen, A., Koh, J., Ploennigs, J., Agarwal, Y., Bergés, M., Culler, D., Gupta, R. K., Kjærgaard, M. B., Srivastava, M., & Whitehouse, K. (2018). Brick: Metadata schema for portable smart building applications. *Applied Energy*, *226*, 1273-1292.

Baracho, R. M. A., Soergel, D., Pereira Junior, M. L., & Henriques, M. A. (2019). A proposal for developing a comprehensive ontology for Smart Cities / Smart Buildings / Smart Life. *Proceedings of the 10th International Multi-Conference on Complexity, Informatics and Cybernetics* (pp. 110-115).

Barbosa dos Santos, M. L. (2022). The "so-called" UGC: An updated definition of user-generated content in the age of social media. *Online Information Review, (46)*1, 95-113.

Barton, A., Duncan, W., Toyoshima, F., & Ethier, J.F. (2018). First steps towards an ontology of belief. *Proceedings of the Joint Ontology Workshops 2018 in the 10th International Conference on Formal Ontology in Information Systems* (pp. 1-5).

Basile, P., Caputo, A., & Semeraro, G. (2014). An enhanced Lesk word sense disambiguation algorithm through a distributional semantic model. *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1591-1600). Association for Computational Linguistics.

Bassi, A., Bauer, M., Fiedler, M., Kramp, T., Van Kranenburg, R., Lange, S., & Meissner, S. (2016). *Enabling things to talk*. Springer.

Beliga, S., Mestrovic, A., & Martincic-Ipsic, S. (2015). An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences, 39*(1), 1-20.

Best, J. (1990). *Threatened children: Rhetoric and concern about child-victims*. University of Chicago Press.

Best, J. (1995). Typification and social problems construction. In: J. Best (Ed.), *Images of issues: Typifying contemporary social problems* (pp. 1-10). Routledge.

Best, J. (2017). *Social problems*. W. W. Norton and & Company.

Blundell, J., Higgens, J., & Middlemiss, N. (1982). *Function in English*. Oxford University Press.

Böhler-Baedeker, S., & Durant, T. (2015). *SUMP Glossary*. http://www.sump-challenges.eu/sites/www.sump-challenges.eu/files/3_ch4llenge_sump_glossary.pdf

Bouzidi, R., De Nicola, A., Nader, F., & Chalal, R. (2019). OntoGamif: A modular ontology for integrated gamification. *Applied Ontology, 14*(3), 215-249.

Brewster, C., Alani, H., Dasmahapatra, S., & Wilks, Y. (2004). Data driven ontology evaluation. *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 641-644).

Bundesregierung (2018). *Government report on wellbeing in Germany*. http://www.gut-leben-in-deutschland.de/downloads/Government-Report-on-Wellbeing-in-Germany.pdf

Cantone, D., Longo, C. F., Nicolosi-Asmundo, M., Santamaria, D. F., & Santoro, C. (2019). Towards an ontology-based framework for a behavior-oriented integration of the IoT. *Proceedings of the 20th Workshop From Objects to Agents* (pp. 119-126).

Cantone, D., Longo, C. F., Nicolosi-Asmundo, M., Santamaria, D. F., & Santoro, C. (2020). Ontological Smart Contracts in OASIS: Ontology for Agents, Systems, and Integration of Services. In: D. Camacho, D. Rosaci, G. M. L. Sarné, & M. Versaci (Eds.), *Intelligent distributed computing XIV*, Studies in Computational Intelligence, vol. 1026 (pp. 237-247). Springer.

Ceusters, W., & Smith, B. (2010). Foundations for a realist ontology of mental disease. *Journal of Biomedical Semantics, 1*(10), 1-23.

Chehade, S., Matta, N., Pothin, J. B., & Cogranne, R. (2020). Handling effective communication to support awareness in rescue operations. *Journal of Contingencies and Crisis Management, 28*, 307-323.

Chimalakonda, S., & Nori, K. (2020). An ontology based modeling framework for design of educational technologies. *Smart Learning Environments, 7*(28), 1-24.

Chomsky, N. A. (1986). *Knowledge of language: Its nature, origin, and use*. Praeger.

Christensen, H. S. (2016). All the same? Examining the link between three kinds of political dissatisfaction and protest. *Comparative European Politics, 14*(6), 781-801.

Chun, S., Jung, J., Jin, X., Seo, S., & Lee, K. H. (2020). Designing an integrated knowledge graph for smart energy services. *The Journal of Supercomputing, 76*, 8058-8085.

Colla, D., Mensa, E., & Radicioni, D. P. (2020). LessLex: Linking multilingual embeddings to sense representations of LEXical items. *Computational Linguistics 46*(2), 289-333.

Doran, D., Gokhale, S., & Dagnino, A. (2013). Human sensing for smart cities. *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1323-1330). ACM.

Doran, D., Severin, K., Gokhale, S., & Dagnino, A. (2016). Social media enabled human sensing for smart cities. *AI Communications, 29*, 57-75.

Eitzen, S., Baca Zinn, M., & Eitzen Smith, K. (2014). *Social problems*, 13th edition. Pearson.

Elmhadhbi, L., Karray, M. H., & Archimède, B. (2019). A modular ontology for semantically enhanced interoperability in operational disaster response. *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management* (pp. 1021-1029).

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors 37*(1), 32-64.

Endsley, M. R. (2000). Theoretical underpinnings of situation awareness: A critical review. In: M. R. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 3-32). Lawrence Erlbaum Associates.

Espinoza-Arias, P., Poveda-Villalón, M., García-Castro, R., & Corcho, O. (2019). Ontological representation of smart city data: From devices to cities. *Applied Sciences, 9*(1), 1-23.

Eurostat (2017). *Final report of the Expert Group on Quality of Life Indicators*. Publications Office of the European Union.

Falco, E., & Kleinhans, R. (2018). Digital participatory platforms for coproduction in urban development: A systematic review. *International Journal of E-Planning Research, 7*(3), 52-79.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Fernández López, M., Gómez-Pérez, A., & Juristo, N. (1997). METHONTOLOGY: From ontological art towards ontological engineering. *Proceedings of the AAAI-97 Symposium on Ontological Engineering* (pp. 33-40).

Fernández López, M., Gómez-Pérez, A., Pazos Sierra, J., & Pazos Sierra, A. (1999). Building a chemical ontology using Methontology and the Ontology Design Environment. *IEEE Intelligent Systems and their Applications, 14*(1), 37-46.

Fernández-Martínez, N. J., & Periñán-Pascual, C. (2021a). LORE: a model for the detection of fine-grained locative references in tweets. *Onomázein 52*, 195-225.

Fernández-Martínez, N. J., & Periñán-Pascual, C. (2021b). nLORE: A linguistically rich deep-learning system for locative reference extraction in tweets. In: E. Bashir, & M. Luštrek (Eds.), *Intelligent Environments 2021. Workshop Proceedings of the 17th International Conference on Intelligent Environments* (pp. 243-254). IOS Press.

Ganitkevitch, J., Van Durme, B., & Callison-Burch, C. (2013). PPDB: The Paraphrase Database. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 758-764). Association for Computational Linguistics.

Gaur, M., Shekarpour, S., Gyrard, A., & Sheth, A. (2019) Empathi: An ontology for emergency managing and planning about hazard crisis. *Proceedings of the 2019 IEEE 13th International Conference on Semantic Computing* (pp. 396-403).

Giffinger, R., Fertner, C., Kramar, H., Kalasek, R., Pichler-Milanovic, N., & Meijers, E. (2007). *Smart cities - Ranking of European medium-sized cities*. http://www.smart-cities.eu/download/smart_cities_final_report.pdf

Givón, T. (1993). *English grammar: A function-based introduction*. John Benjamins.

Goswami, A. & Kumar, A. (2016). A survey of event detection techniques in online social networks. *Social Network Analysis and Mining, 6*(1), 1-25.

Götz, K. (2014). Traffic mobility. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 6705-6710). Springer.

Govada, S. S., Spruijt, W., & Rodgers, T. (2017). Smart city concept and framework. In: T. M. V. Kumar (Ed.), *Smart economy in smart cities. Advances in 21st century human settlements* (pp. 187-198). Springer.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Gu, J., Gao, B., Chen, Y., Jiang, L., Gao, Z., Ma, X., Ma, Y., & Woo, W. L. (2017). Wearable social sensing and its application in anxiety assessment. *Proceedings of the 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (pp. 305-308). IEEE.

Haddow, G. D., Bullock, J. A., & Coppola, D. P. (2011). *Introduction to emergency management* (4th ed.). Elsevier.

Hastings, J., Ceusters, W., Jensen, M., Mulligan, K., & Smith, B. (2012). Representing mental functioning: Ontologies for mental health and disease. *Proceedings of the Third International Conference on Biomedical Ontology* (pp. 1-5).

Helliwell, J. F., Layard, R., Sachs, J. D., & De Neve, J.-E. (Eds.) (2020). *World happiness report 2020*. Sustainable Development Solutions Network.

Hellmund, T., Hertweck, P., Hilbring, D., Mossgraber, J., Alexandrakis, G., Pouli, P., Siatou, A., & Padeletti, G. (2018) Introducing the HERACLES Ontology—Semantics for Cultural Heritage Management. *Heritage, 1*(2), 377-391.

Hitzler, P., Gangemi, A., Janowicz, K., Krisnadhi, A., & Presutti, V. (Eds.) (2016). *Ontology engineering with ontology design patterns: Foundations and applications*, Studies on the Semantic Web, vol. 25. IOS Press.

Howes, R., & Robinson, H. (2005). *Infrastructure for the built environment: Global procurement strategies*. Elsevier.

Hu, X., Chen, Q., & Du, M. (2020). Ontology-based multi-sensor information integration model for urban gardens and green spaces. *Proceedings of the 2020 International Conference on Green Development and Environmental Science and Technology*, IOP Conference Series: Earth and Environmental Science, vol. 615 (pp. 1-8). IOP Publishing.

Hutchison, W., Bedford, N., & Bedford, S. (2011). Ukraine's global strategy in the post-crisis economy: Developing an intelligent nation to achieve a competitive advantage. *Innovative Marketing, 7*(1), 46-53.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys, 47*(4), 67:1-67:38.

Istituto Nazionale di Statistica (2020). *BES 2020: Il benessere equo e sostenibile in Italia*. https://www.istat.it/it/files//2021/03/BES_2020.pdf.

Janowicz, K., Haller, A., Cox, S. J. D., Le Phuoc, D., & Lefrançois, M. (2018). SOSA: A lightweight ontology for sensors, observations, samples, and actuators. *Journal of Web Semantics, 56*, 1-10.

Jashapara, A. (2005). *Knowledge management: An integrated approach*. Prentice Hall.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1746-1751). Association for Computational Linguistics.

Komninos, N., Bratsas, C., Kakderi, C., & Tsarchopoulos, P. (2015). Smart city ontologies: Improving the effectiveness of smart city applications. *Journal of Smart Cities, 1*(1), 31-46.

Komninos, N., Panori, A., & Kakderi, C. (2020). The Smart City Ontology 2.0. URENIO Research Discussion Papers. https://www.urenio.org/2020/12/24/smart-city-ontology-2-0/

Kott, J., & Kott, M. (2019) Generic ontology of energy consumption households. *Energies, 12*(19), 3712, 1-19.

Krisnadhi, A. & Hitzler, P. (2016). Modeling with ontology design patterns: Chess games as a worked example. In: P. Hitzler, A. Gangemi, K. Janowicz, A. Krisnadhi, & V. Presutti (Eds.), *Ontology engineering with ontology design patterns: Foundations and applications*, Studies on the Semantic Web, vol. 25 (pp. 3-21). IOS Press.

Krötzsch, M., Simančík, F., & Horrocks, I. (2014). A Description Logic primer. In: J. Lehmann, & J. Völker (Eds.), *Perspectives on Ontology Learning* (pp. 3-19). IOS Press.

Kurte, K., Potnis, A., & Durbha, S. (2019). Semantics-enabled spatio-temporal modeling of earth observation data: An application to flood monitoring. *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities* (pp. 41-50).

Letukas, L. (2014). *Primetime pundits: How cable news covers social issues*. Lexington Books.

Lindell, M. K., Prater, C. S., & Perry R. W. (2006). *Introduction to emergency management*. Wiley Pathways.

Loseke, D. R. (2017). *Thinking about social problems: An introduction to constructionist perspectives*. Routledge.

Madan, A., Cebrián, M., Lazer, D., & Pentland, A. (2010). Social sensing for epidemiological behavior change. *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 291-300). ACM.

Madukwe, K. J., Gao, X., & Xue, B. (2021). What emotion is hate? Incorporating emotion information into the hate speech detection task. In: D. N. Pham, T. Theeramunkong, G. Governatori, & F. Liu (Eds.), *PRICAI 2021: Trends in Artificial Intelligence*, Lecture Notes in Computer Science, vol. 13032 (pp. 273-286). Springer.

Marginean, I. (2014). Quality of Life Diagnosis (QoLD). In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 5333-5339). Springer.

Maru, M., Scozzafava, F., Martelli, F., & Navigli, R. (2019). SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3532-3538). Association for Computational Linguistics.

McCain, K. (2016). *The nature of scientific knowledge*. Springer.

McGuinness, D. L. (2003). Ontologies come of age. In D. Fensel, W. Wahlster, H. Lieberman, & J. Hendler (Eds.), *Spinning the semantic web: Bringing the world wide web to its full potential* (pp. 171-194). MIT Press.

McShane, M. (2017). Natural Language Understanding (NLU, not NLP) in cognitive systems. *AI Magazine, 38*(4), 43-56.

Moreira, J. L. R., Pires, L., Sinderen, M., Daniele, L., & Girod-Genet, M. (2020). SAREF4health: Towards IoT standard-based ontology-driven cardiac e-health systems. *Applied Ontology, 15*(3), 385-410.

Musto, C., Semeraro, G., Lops, P., & De Gemmis, M. (2015). CrowdPulse: A framework for real-time semantic analysis of social streams. *Information Systems, 54*, 127-146.

Nalchigar, S., & Fox, M. S. (2018). Achieving interoperability of Smart City data: An analysis of 311 data. *Journal of Smart Cities, 3*(1), 1-13.

Nasim, Z., & Khan, I. (2018). Solving poverty using ontology. *Proceedings of the 10th International Conference on Knowledge Engineering and Ontology Development* (pp. 271-278).

Neuman, M. (2005). Infrastructure. In R. W. Caves (Ed.), *Encyclopedia of the city* (pp. 385-389). Routledge.

OECD (2020). *How's life? 2020: Measuring well-being*. OECD Publishing. https://www.oecd-ilibrary.org/economics/how-s-life/volume-/issue-_9870c393-en

Office for National Statistics (2019). *Measuring national well-being in the UK: International comparisons 2019*. https://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/articles/measuringnationalwellbeing/internationalcomparisons2019

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Panori, A., Kakderi, C., & Tsarchopoulos, P. (2019). Designing the ontology of a Smart City application for measuring multidimensional urban poverty. *Journal of the Knowledge Economy, 10*, 921-940.

Parrillo, V. N. (Ed.). (2008). *Encyclopedia of social problems*. Sage.

Peng, C., & Goswami, P. (2019). Meaningful integration of data from heterogeneous health services and home environment based on ontology. *Sensors, 19*(8), 1747, 1-19.

Peterwagner, R. (2005). *What is the matter with communicative competence?*. LIT.

Pluwak, A. (2016). Towards the application of speech act theory to opinion mining. *Cognitive Studies| Études Cognitives, 16*, 33-44.

Polanyi, L., & Zaenen, A. (2004). Contextual valence shifters. *Working Notes of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications* (pp. 106-111). The AAAI Press.

Qamar, T., Bawany, N. Z., Javed, S., & Amber, S. (2019). Smart City Services Ontology (SCSO): Semantic modeling of smart city applications. *Proceedings of the 7th International Conference on Digital Information Processing and Communications* (pp. 52-56).

Quirk, R., Greenbaum, S, Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Ramaprasad, A., Sánchez-Ortiz, A., & Syn, T. (2017). A unified definition of a Smart City. *Proceedings of the 16th IFIP WG 8.5 International Conference* (pp. 13-24). Springer.

Rani, M., Alekh, S., Bhardwaj, A., Gupta, A., & Vyas, O. P. **(**2016). Ontology-based classification and analysis of non-emergency smart-city events. *Proceedings of the 2016 International Conference on Computational Techniques in Information and Communication Technologies* (pp. 509-514).

Rose, D., & Schaffer, J. (2013). Knowledge entails dispositional belief. *Philosophical Studies 166*, 19-50.

Saha, T., Jayashree, S. R., & Saha, S. (2020). BERT-Caps: A transformer-based capsule network for tweet act classification. *IEEE Transactions on Computational Social Systems, 7*(5), 1168-1179.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 919-931.

Sawesi, S. (2018). *An ontology for formal representation of medication adherence-related knowledge: Case study in breast cancer*, PhD thesis. Indiana University.

Schwab, A. K., Eschelbach, K., & Brower, D. J. (2007). *Hazard mitigation and preparedness*. John Wiley & Sons.

Searle, J. R. (1969). *Speech acts*. Cambridge University Press.

Seccombe, K. & Kornblum, W. (2020). *Social problems*, 16th edition. Pearson.

Sheridan, P., Onsjö, M., & Hastings, J. (2019). The literary theme ontology for media annotation and information retrieval. *Proceedings of the Joint Ontology Workshops 2019* (pp. 1-12).

Sirgy, M. J. (2018). The psychology of material well-being. *Applied Research Quality Life, 13*, 273-301.

Smith, B., & Ceusters, W. (2015). Aboutness: Towards foundations for the Information Artifact Ontology. *Proceedings of the Sixth International Conference on Biomedical Ontology* (pp. 47-51).

Soergel, D. (2008). Digital libraries and knowledge organization systems. In S. Kruk (Ed.), *Semantic digital libraries* (pp. 9-39). Springer.

Soergel, D., Baracho, R. M. A., & Mullarkey, M. (2020). Toward a comprehensive Smart Ecosystem Ontology - Smart Cities, Smart Buildings, Smart Life. *Journal of Systemics, Cybernetics and Informatics, 18*(2), 25-36.

Sperber, D., & Wilson, D. (2015). Beyond speaker's meaning. *Croatian Journal of Philosophy, 15*(44), 117-149.

Spoladore, D. Mahroo, A. Trombetta, & Sacco, M. (2019). ComfOnt: A semantic framework for indoor comfort and energy saving in smart homes. *Electronics, 8*, 1449, 1-21.

Syzdykbayev, M., Hajari, H., & Karimi, H. A. (2019). An ontology for collaborative navigation among autonomous cars, drivers, and pedestrians in smart cities. *Proceedings of the 4th International Conference on Smart and Sustainable Technologies* (pp. 1-6).

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind*. Oxford University Press.

Tiwari, S., & Abraham, A. (2020). Semantic assessment of smart healthcare ontology. *International Journal of Web Information Systems, 16*(4), 475-491.

Toyoshima, F., Barton, A., & Grenier, O. (2020). Foundations for an ontology of belief, desire and intention. *Proceedings of the 11th International Conference on Formal Ontology in Information Systems* (pp. 140-154). IOS Press.

Valkenburg, R., Den Ouden, E., & Schreurs, M. A. (2016). Designing a smart society: From smart cities to smart societies. In B. Salmelin (Ed.), *Open Innovation 2.0 yearbook 2016* (pp. 87-92). European Commission.

Viktorović, M., Yang, D., & Vries, B. D. (2020). Connected Traffic Data Ontology (CTDO) for intelligent urban traffic systems focused on connected (semi) autonomous vehicles. *Sensors, 20*(10), 2961, 1-14.

Vosoughi, S. & Roy, D. (2016). Tweet Acts: A speech act classifier for Twitter. *Proceedings of the Tenth International AAAI Conference on Web and Social Media* (pp. 711-714).

Wang, D., Szymanski, B. K., Abdelzaher, T., Ji, H., & Kaplan, L. (2019). The age of social sensing. *IEEE Computer 52*(1), 36-45.

Wei, L., Du, H., Mahesar, Q.A., Al Ammari, K., Magee, D., Clarke, B., Dimitrova, V., Gunn, D., Entwisle, D., Reeves, H., & Cohn, A. G. (2020). A decision support system for urban infrastructure inter-asset management employing domain ontologies and qualitative uncertainty-based reasoning. *Expert Systems with Applications 158*, 113461, 1-24.

Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 6381-6387).

Wilkins, D. A. (1976). *Notional syllabuses*. Oxford University Press.

Wyrwoll, C. (2014). *Social Media: Fundamentals, Models, and Ranking of User-Generated Content*. Doctoral Thesis, Universität Hamburg, Springer Vieweg.

Xu, Z., Fu, Y., Chen, X., Rao, Y., Xie, H., Wang, F. L., & Peng, Y. (2018). Sentiment classification via supplementary information modeling. In: Y. Cai, Y. Ishikawa, & J. Xu (Eds.) *Web and Big Data. Second International Joint Conference, APWeb-WAIM 2018*, Lecture Notes in Computer Science, vol. 10987 (pp. 54-62). Springer.

Xu, Z., Mei, L., Choo, K.-K. R., Lv, Z, Hu, C., Luo, X., & Liu, Y. (2018). Mobile crowd sensing of human-like intelligence using social sensors: A survey. *Neurocomputing 279*, 3-10.

Zhang, R., Gao, D., & Li, W. (2011). What are tweeters doing: Recognising speech acts in Twitter. In *Proceedings of the 5th AAAI Conference on Analyzing Microtext* (pp. 86-91).

Zhang, C., & Xu, H. (2009). Using Citation-KNN for automatic keyword assignment. *Proceedings of the 2009 International Conference on Electronic Commerce and Business Intelligence* (pp. 131-134).