Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Ciência da Computação

Doctoral Thesis

Ontology-driven Urban Issues Identification from Social Media

Maxwell Guimarães de Oliveira

Campina Grande, Paraíba, Brazil
2016

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Ciência da Computação

Tese de Doutorado

Identificação de Problemas Urbanos em Mídias Sociais dirigida por Ontologia

(*title in Portuguese*)

Maxwell Guimarães de Oliveira

Campina Grande, Paraíba, Brasil
2016

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Ciência da Computação

Ontology-driven Urban Issues Identification from Social Media

Maxwell Guimarães de Oliveira

Doctoral thesis submitted to the Computer Science Post Graduate Program from the Electrical Engineering Centre at Federal University of Campina Grande, as a partial requirement for obtaining a PhD degree in Computer Science.

Concentration Area: Computer Science

Research Line: Computing Systems

Supervisors:

Cláudio de Souza Baptista, Ph.D

Cláudio Elízio Calazans Campelo, Ph.D

Campina Grande, Paraíba, Brazil
December, 2016

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

O48o Oliveira, Maxwell Guimarães de.

Ontology-driven urban issues identification from social media / Maxwell Guimarães de Oliveira. – Campina Grande, 2016.

170 f. : il. color.

Tese (Doutorado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2017.

"Orientação: Prof. Dr. Cláudio de Souza Baptista; Coorientação: Prof. Dr. Cláudio Elízio Calazans Campelo".

1. Crowdsourcing. 2. Geoparsing. 3. Ontologia. 4. Mídia Social. 5. Problemas Urbanos. I. Baptista, Cláudio de Souza. II. Campelo, Cláudio Elízio Calazans. III. Universidade Federal de Campina Grande, Campina Grande (PB). IV. Título.

CDU 004.822(043)

Abstract

The cities worldwide face with many issues directly related to the urban space, especially in the infrastructure aspects. Most of these urban issues generally affect the life of both resident and visitant people. For example, people can report a car parked on a footpath which is forcing pedestrians to walk on the road or a huge pothole that is causing traffic congestion. Besides being related to the urban space, urban issues generally demand actions from city authorities. There are many Location-Based Social Networks (LBSN) in the smart cities domain worldwide where people complain about urban issues in a structured way and local authorities are aware to fix them. With the advent of social networks such as Facebook and Twitter, people tend to complain in an unstructured, sparse and unpredictable way, being difficult to identify urban issues eventually reported. Social media data, especially Twitter messages, photos, and check-ins, have played an important role in the smart cities. A key problem is the challenge in identifying specific and relevant conversations on processing the noisy crowdsourced data. In this context, this research investigates computational methods in order to provide automated identification of urban issues shared in social media streams. Most related work rely on classifiers based on machine learning techniques such as Support Vector Machines (SVM), Naïve Bayes and Decision Trees; and face problems concerning semantic knowledge representation, human readability and inference capability. Aiming at overcoming this semantic gap, this research investigates the ontology-driven Information Extraction (IE) from the perspective of urban issues; as such issues can be semantically linked in LBSN platforms. Therefore, this work proposes an Urban Issues Domain Ontology (UIDO) to enable the identification and classification of urban issues in an automated approach that focuses mainly on the thematic and geographical facets. Experimental evaluation demonstrates the proposed approach performance is competitive with most commonly used machine learning algorithms applied for that particular domain.

Keywords: Crowdsourcing, Geoparsing, Ontology, Social Media, Urban Issues.

Resumo

As cidades em todo o mundo enfrentam muitos problemas diretamente relacionados ao espaço urbano, especialmente nos aspectos de infraestrutura. A maioria desses problemas urbanos geralmente afeta a vida de residentes e visitantes. Por exemplo, as pessoas podem relatar um carro estacionado em uma calçada que está forçando os pedestres a andar na via, ou um enorme buraco que está causando congestionamento. Além de estarem relacionados com o espaço urbano, os problemas urbanos geralmente demandam ações das autoridades municipais. Existem diversas Redes Sociais Baseadas em Localização (*LBSN*, em inglês) no domínio das cidades inteligentes em todo o mundo, onde as pessoas relatam problemas urbanos de forma estruturada e as autoridades locais tomam conhecimento para então solucioná-los. Com o advento das redes sociais como Facebook e Twitter, as pessoas tendem a reclamar de forma não estruturada, esparsa e imprevisível, sendo difícil identificar problemas urbanos eventualmente relatados. Dados de mídia social, especialmente mensagens do Twitter, fotos e check-ins, tem desempenhado um papel importante nas cidades inteligentes. Um problema chave é o desafio de identificar conversas específicas e relevantes ao processar dados *crowdsourcing* ruidosos. Neste contexto, esta pesquisa investiga métodos computacionais a fim de fornecer uma identificação automatizada de problemas urbanos compartilhados em mídias sociais. A maioria dos trabalhos relacionados depende de classificadores baseados em técnicas de aprendizado de máquina, como SVM, *Naïve Bayes* e Árvores de Decisão; e enfrentam problemas relacionados à representação do conhecimento semântico, legibilidade humana e capacidade de inferência. Com o objetivo de superar essa lacuna semântica, esta pesquisa investiga a Extração de Informação baseada em ontologias, a partir da perspectiva de problemas urbanos, uma vez que tais problemas podem ser semanticamente interligados em plataformas LBSN. Dessa forma, este trabalho propõe uma ontologia no domínio de Problemas Urbanos (UIDO) para viabilizar a identificação e classificação dos problemas urbanos em uma abordagem automatizada que foca principalmente nas facetas temática e geográfica. Uma avaliação experimental demonstra que o desempenho da abordagem proposta é competitivo com os algoritmos de aprendizado de máquina mais utilizados, quando aplicados a este domínio em particular.

Palavras-chave: Crowdsourcing, Geoparsing, Ontologia, Mídia Social, Problemas Urbanos.

Acknowledgments

I am first grateful to God for lighten my ways, giving me health and strength to face this long four-year journey.

I would like to thank my parents, *Seu Zé* (*Mr. José de Oliveira*) and *Dona Lena* (*Ms. Edilene Guimarães*) for all the investment, effort, dedication and support throughout my life, without which I would not be able to reach this achievement.

I would like to thank my lovely wife *Selma*, for her presence, affection and continuous support; for giving me the blessing of being the father of my beloved son *Lucas*; and also for understanding at times when I could not make present myself because of this work. Thank my brother *Everton*, my cousin *Josi*, and all from my family for all the joys in family moments.

I would like to thank my supervisors, *Dr. Cláudio Baptista* and *Dr. Cláudio Campelo*, for all the support, motivation and valuable teachings not only for this document, but also for my professional and personal life. Thank my co-supervisor *Dr. Michela Bertolotto*, for all the support and knowledge transmitted both face-to-face and aloof.

I would like to thank my viva examiners, represented by *Dr. Fabio Gomes, Dr. Joseana Fechine, Dr. Leandro Balby* and *Dr. Renato Fileto*, for the valuable contributions in this document and research.

I would like to thank the researchers who I had the opportunity to interact with during the doctorate and who contributed in some way to my formation: *Dr. Geraldo Braz, Dr. Peter Mooney, Dr. Muki Haklay, Dr. Padraig Corcoran, Dr. Sergio Di Martino, Dr. Pasquale Di Giovanni, Dr. Musfira Jilani and Msc. Laura Di Rocco*.

I would like to thank the colleagues from LSI/UFCG (Luiz, Davi, Amilton, Ana Gabrielle, Ana Paula, Anderson, André, Brunna, Caio, Damião, Daniel, Francisco, Gabriel, Hugo, Igor, Jaindson, Júlio, Nathaniel, Ruan, Tiago, Yuri and Wener) for the support in this research, for the exchange of ideas, incentives, experiences and discussions of random subjects, always accompanied by good cups of coffee.

I would like to thank my longtime friends in Campina Grande: Bruno Dias, Danilo Abreu, Daniel Fireman, Rodrigo Noel, Marco Rosner, Adriano Santos, Elthon Oliveira, Myslane Farias, Aninha, Henrique Farias, Val Couto, Viviano Moura, Carlinda Aragão, Tiago and Lucélia Barbosa, for the relaxation moments, lunches, dinners, junk food, coffees and craft beers.

I would like to thank my friends in Dublin: Renata Barros, Carol Correia, Fabiola Neto, Dario Borrego, Michele Menis, Camila Blanco, Isadora Henrichs, Thayse Passos, Fernando Salles and Warren Byrne, for all the pints of Guinness, Hop House 13 and many others; for the fun trips on Dublin Bus, as well as great meetings always in companion with good beers, coffees and limoncello.

I would like to thank my UFCG colleagues: Selma Torquato, Fabiana França, Teresa Péret (*Cris*), Maria do Socorro Pedrosa (*Nana*), Geralda Alves, Ianna Kobayashi, Tarcísio Araujo (*Tioba*), Fernando Careca, and Edivandro Barros, for all the support, fellowship and friendship.

I would like to thank the "Wind-FM Classic Rock Station", for the continuous partnership by providing the best of Rock n' Roll music and making work moments more excited.

Last, but not least, I would like to thank UFCG for granting me leave away from the job for exclusive dedication to the doctorate; I would like to thank CNPq for providing the financial support that enabled the improvement of studies abroad; I would like to thank LSI/UFCG and SCI/UCD for providing infrastructure to carry out the work; and I would like to thank COPIN for the bureaucratic support in what was needed.

Finally, I would like to thank everyone who has contributed in some way to this achievement, to whom I apologize for not remembering and mentioning directly.

Agradecimentos

Primeiramente, agradeço a Deus, por iluminar os meus caminhos, me dando saúde e forças para enfrentar essa longa jornada de quatro anos.

Agradeço aos meus pais, *Seu Zé* (*José de Oliveira*) e *Dona Lena* (*Edilene Guimarães*) por todo o investimento, esforço, dedicação e suporte ao longo da minha vida, sem os quais eu não conseguiria mais esta realização.

À minha querida esposa *Selma*, pela presença, pelo carinho, pelo apoio contínuo, por me proporcionar a benção de ser pai do meu amado filho *Lucas*, e também pela compreensão nos momentos em que não pude me fazer presente em virtude deste trabalho. Ao meu irmão *Everton*, à minha prima *Josi*, e aos demais familiares queridos, por todas as alegrias dos momentos em família.

Aos meus orientadores, professores *Dr. Cláudio Baptista* e *Dr. Cláudio Campelo*, por todo o suporte, motivação e ensinamentos valiosos não somente para este documento, como para minha vida profissional e pessoal. À minha co-orientadora, professora *Dra. Michela Bertolotto*, por todo o suporte e conhecimento transmitido, tanto presencial quanto à distância.

À banca examinadora, representada pelos professores *Dr. Fabio Gomes*, *Dra. Joseana Fechine*, *Dr. Leandro Balby* e *Dr. Renato Fileto*, pelas valiosas contribuições nesta pesquisa e neste documento.

Aos pesquisadores que tive a oportunidade de interagir no doutorado e que contribuíram de alguma forma em minha formação: *Dr. Geraldo Braz*, *Dr. Peter Mooney*, *Dr. Muki Haklay*, *Dr. Padraig Corcoran*, *Dr. Sergio Di Martino*, *Dr. Pasquale Di Giovanni*, *Dra. Musfira Jilani e Msc. Laura Di Rocco*.

Aos colegas do LSI/UFCG (Luiz, Davi, Amilton, Ana Gabrielle, Ana Paula, Anderson, André, Brunna, Caio, Damião, Daniel, Francisco, Gabriel, Hugo, Igor, Jaindson, Júlio, Nathaniel, Ruan, Tiago, Yuri e Wener) pelo apoio nesta pesquisa, troca de ideias,

incentivos, experiências e discussões de assuntos aleatórios sempre regados a boas xícaras de café.

Aos meus amigos de longa data em Campina Grande: Bruno Dias, Danilo Abreu, Daniel Fireman, Rodrigo Noel, Marco Rosner, Adriano Santos, Elthon Oliveira, Myslane Farias, Aninha, Henrique Farias, Val Couto, Viviano Moura, Carlinda Aragão, Tiago e Lucélia Barbosa, pelos momentos de descontração, almoços, jantares, gordices, cafés e cervejas artesanais.

Aos amigos que fiz em Dublin: Renata Barros, Carol Correia, Fabíola Neto, Dario Borrego, Michele Menis, Camila Blanco, Isadora Henrichs, Thayse Passos, Fernando Salles e Warren Byrne, por todas as pints de Guinness, Hop House 13 e tantas outras; pelas divertidas viagens de Dublin Bus, além das ótimas reuniões sempre regadas a boas cervejas, bons cafés e *limoncello*.

Aos colegas da UFCG: Selma Torquato, Fabiana França, Teresa Péret (*Cris*), Maria do Socorro Pedrosa (*Nana*), Geralda Alves, Ianna Kobayashi, Tarcísio Araujo (*Tioba*), Fernando Careca e Edivandro Barros, por todo o apoio, companheirismo e amizade.

Agradeço à "*Wind-FM Classic Rock Station*", pela parceria contínua, fornecendo o melhor do *Rock n' Roll* e deixando mais animado os momentos de trabalho.

Por fim, mas não menos importante, à UFCG, por me conceder afastamento para dedicação exclusiva ao doutorado, ao CNPq, pelo apoio financeiro que viabilizou o aperfeiçoamento dos estudos no exterior, ao LSI/UFCG e à SCI/UCD, pela infraestrutura, e à COPIN, pelo apoio burocrático no que foi necessário.

Enfim, gostaria de agradecer a todos que contribuíram de alguma forma nesta conquista, a quem me desculpo desde já por não recordar e mencionar diretamente.

I dedicate this thesis to my grandfather **Euclides Alves de Morais** (*in memorian*), a sage man who realized my educational potential when I still was a child; and who left his legacy of wisdom and character even departing early from this world.

Dedico esta tese ao meu avô **Euclides Alves de Morais** (*in memorian*), um homem sábio, que percebeu meu potencial para os estudos ainda criança, e que, mesmo partindo cedo deste mundo, deixou seu legado de sabedoria e caráter.

Table of Contents

List of Abbreviations

AGI	Ambient Geospatial Information
API	Application Programming Interface
AUC	Area under an ROC curve
CRF	Conditional Random Field
GIR	Geographic Information Retrieval
GLoD	Geographic Level of Detail
GMT	Greenwich Mean Time
GPS	Global Positioning System
IE	Information Extraction
IR	Information Retrieval
JSON	JavaScript Object Notation
KNN	k-Nearest Neighbors
LBS	Location-Based Services
LBSN	Location-Based Social Network
NER	Named-Entity Recognition
NLP	Natural Language Processing
OSM	OpenStreetMap
OWL	Web Ontology Language
POI	Point-Of-Interest
RDBMS	Relational Database Management System
SVM	Support Vector Machines
UIDO	Urban Issues Domain Ontology
UK	United Kingdom
URL	Uniform Resource Locator
VGI	Volunteered Geographic Information
XML	Extensible Markup Language

List of Figures

List of Tables

List of Listings

Chapter 1

Introduction

Smart cities is a key concept (Kitchin, 2014; Caragliu et al., 2011) that has become quite popular in the last years. The term "smart" refers to the ability of solving urban issues. Nowadays, there are a number of urban sensors that may be explored to produce a well detailed understanding of the city dynamics, enabling the discovering of urban issues. Citizens have become true human sensors around urban areas with the advent of mobile devices and the development of social media networks.

The phenomenon widely known as Crowdsourcing (Surowiecki, 2005) involves online communities that combine the efforts of self-identified volunteers to produce valuable crowdsourced data in various different domains. Such phenomenon contributed to the emergence of Location-Based Social Networks (LBSNs) for the smart cities domain. This kind of LBSN comprises crowdsourcing environments that have enabled citizens to share valuable semantic information regarding urban neighborhoods, including spatial and temporal aspects.

Crowdsourced data allow the exploration and development of a new kind of urban science that needs an interdisciplinary approach (Thrift, 2014). Such interdisciplinarity is supported by urban computing, an emerging research field where computer science meets city-related fields in the context of urban spaces (Zheng et al., 2014a). To date, the information shared through LBSNs is appreciated by city councils, administrative authorities that have made efforts to solve most of the reported issues. Therefore, such kind

of social network plays an important role in enabling an easier communication between the citizens and the government (Crooks et al., 2015).

Although government response regarding the reported issues has been initially thought as a key motivation for citizens to become active users and producers of crowdsourced data, typically only a few users provide a significant amount of information. This phenomenon is visible in terms of Volunteered Geographical Information (VGI), where many areas around the world are mapped only by a few users (Haklay and Weber, 2008). On the other hand, the number of active users in popular social media networks such as Facebook[1] and Twitter[2] has increased considerably.

Social media data, especially the geotagged Twitter messages (*tweets*), photos, and check-ins, have played an important role in revealing individuals' life patterns (Ye et al., 2011), lifestyles (Yuan et al., 2013), behavior patterns (Wakamiya et al., 2012) and cities' dynamics (Cranshaw et al., 2012). Besides sharing their personal activities, such active users often also share issues about their neighborhoods or city surroundings. From the user's point of view, it is easier to complain, suggest or even comment on a platform they often use and that has a huge audience such as Twitter (Gelernter and Mushegian, 2011). Most people tend to concentrate their online conversation in a small number of preferred systems, instead of constantly discovering, signing up and learning to use new systems as they are launched. Furthermore, three location contexts can be assigned to a tweet: the geocoded, the user home and the mentioned location; where each tweet location context relates to a geographical location that can be explored in many different ways. In this context, this thesis focuses on the problem of automatically extracting urban issues reported on social media streams in order to enrich LBSNs for the smart cities domain with geocoded content.

Urban issues is a term given to reported issues related to the urban infrastructure, such as potholes in a road, rubbish in a forbidden place, and faulty traffic lights, among others. Thus, this thesis relies on the assumption that social media messages can be automatically annotated, playing a role in LBSN enrichment and linking both relevant and up-to-date information in the smart cities domain. Once such link is established, urban issues reported on social media can also be efficiently appreciated by administrative

[1] http://www.facebook.com/
[2] http://www.twitter.com/

authorities in smart cities environments, which are responsible for fixing the problems and providing adequate feedback to the citizens.

While LBSNs for the smart cities domain are driven by a specific vocabulary to be used by the users, popular social networks seem to be a generic forum where people talk about everything in different ways. Discovering specific and relevant conversations about certain types of subjects is a challenge in any kind of text. In the social media context it is even more challenging, since social media messages tend to be short, unstructured, based on informal language and vernacular terms (Bordogna et al., 2012). Those issues, combined with the advent of social media, have enabled the emergence of the social media analytics research field, as traditional Information Retrieval (IR) and Geographic Information Retrieval (GIR) techniques do not work properly with social media messages. Therefore, novel techniques and algorithms that properly deal with social media are required, enabling parsing systems to extract even more relevant information.

Up to the moment, no research has specifically addressed the problem of automated detection of urban issues from social media streams. At first glance, urban issues could be treated as events, so that existing event detection techniques could be applied to perform such an automated extraction of urban issues. There are many proposals on event detection from social media, such as natural hazards (Wang and Stewart, 2015; Imran et al., 2014), traffic (Anantharam et al., 2015), forest fires (Spinsanti and Ostermann, 2013; Abel et al., 2012) and urban events (Xia et al., 2015). Unfortunately, pure event detection techniques seem to be not enough for urban issues identification. While most existing techniques are based on previously known forthcoming events (e.g. analyzing historical time series) or responsive behaviors (e.g. abnormal peaks in a short time span), urban issues are sparse in both time and space, which make such techniques useless for such purpose.

An urban issue report can be naturally seen through three facets: thematic, geographical and temporal. While the geographical facet involves the identification of geographical locations with suitable precision within urban areas and the temporal facet deals with time expressions, the thematic facet involves the identification and classification of these issues. Therefore, the technical problem of urban issue identification from social media can be split into three main and complex sub problems to be addressed. In order to make this thesis feasible, the scope of the research focuses mainly on the thematic and

geographical facets in English texts shared on Twitter, while the temporal facet is minimally addressed.

This thesis investigates the methodologies for complaints identification from social media aiming at developing a proper approach to the automated identification of urban issues. Most proposals from the state-of-the art rely on most commonly used machine learning algorithms to build their classifiers (Jin et al., 2013; Augustine et al., 2012; Yang et al., 2011) and face problems concerning semantic knowledge representation. On the other hand, ontology-driven Information Extraction (IE) has emerged as a powerful approach for extracting spatiotemporal and thematic information in specific domains (Wang and Stewart, 2015). One of the main advantages in using ontologies is the interoperability by humans and the ability of performing inferences, which makes the domain modeling easier and readable by humans. However, the performance of ontology-driven IE and classifiers based on most used machine learning algorithms may vary according to the applied domain. In this context, this research investigates the ontology-driven IE from the perspective of urban issues concerning feasibility and performance in comparison with several machine learning classifiers widely used in the literature.

1.1 Research Questions

Four research questions are addressed in this thesis proposal based on four hypotheses. The context applied on such questions concerns the classification methodology, knowledge learning process, and the thematic and geographical facets. Moreover, it is intended to answer such questions by acquiring and analyzing real data, carrying out case studies and performing statistical analysis. Each of the questions is presented with its respective null (H_0) and alternative hypotheses (H_1).

There are two main questions that concern this thesis: Q1 and Q2. These questions address the classification methodology developed in this work for the automated identification of urban issues from social media data.

Q1: Is ontology-based IE performance competitive with the most commonly used machine learning algorithms in identifying urban problems reported on social networks such as Twitter, with the knowledge learned from LBSN for that particular domain?

$H1_0$: Ontology-driven IE is not competitive with the most commonly used machine learning algorithms with knowledge learned from LBSN for the urban issues domain.

$H1_1$: Ontology-driven IE is competitive with the most commonly used machine learning algorithms with knowledge learned from LBSN for the urban issues domain.

Q2: Is ontology-based IE performance competitive with the most commonly used machine learning algorithms in identifying urban problems reported on social networks such as Twitter, with knowledge learned from manually labeled tweets for that particular domain?

$H2_0$: Ontology-driven IE is not competitive with the most commonly used machine learning algorithms with knowledge learned from manually labeled tweets for the urban issues domain.

$H2_1$: Ontology-driven IE is competitive with the most commonly used machine learning algorithms with knowledge learned from manually labeled tweets for the urban issues domain.

While Q1 focuses on the knowledge learned from LBSN in the urban issues domain, Q2 focuses on the knowledge learned from labeled tweets. The nature of the corpora adopted for the knowledge learning may influence the results because tweets are shorter than LBSN reports and may present labeling errors. Therefore, such questions are answered through case studies that statistically compare the classifiers.

This thesis also addresses two auxiliary research questions (Q3 and Q4) that are related to the facets involved in the automated identification of urban issues in tweets.

Q3: In geocoded tweets, is the geocoded location reliable to replace geographical locations eventually mentioned in a tweet regarding the context of urban areas?

$H3_0$: The geocoded location is not reliable to replace geographical locations eventually mentioned in a tweet regarding the context of urban areas.

$H3_1$: The geocoded location is reliable to replace geographical locations eventually mentioned in a tweet regarding the context of urban areas.

Q4: Is the sentiment polarity of a tweet determinant for the identification of an urban issue being reported?

$H4_0$: The sentiment polarity of a tweet is determinative for the identification of an urban issue.

$H4_1$: The sentiment polarity of a tweet is not determinative for the identification of an urban issue.

While Q3 focuses on the geographical facet of urban issues, Q4 focuses on the thematic facet. Answering Q3 is important because geoparsing an urban location mentioned in a tweet message is challenging, costly and not well accurate. Thus, using the geocoded location from the tweet metadata instead of performing geoparsing in the tweet message would reduce processing costs. The intention is to answer Q3 by analyzing samples of geocoded tweets from two different urban areas which had the tweet messages manually geoparsed to urban locations.

Finally, answering Q4 is important in order to give an indication as to whether sentiment analysis can be useful in the identification of urban issues, as sentiment analysis is both challenging and costly. The intention is to answer Q4 by analyzing a sample of manually labeled tweets concerning urban issues and sentiment polarity.

1.2 Goals

The main goal of this thesis is to develop an approach to the automated identification of urban issues from social media that considers the semantics of such specific domain; that can be easily improved by specialist knowledge; and that presents a competitive performance in comparison with classifiers based on most commonly used machine learning algorithms.

The following specific goals were defined in order to reach the main goal:

- to research techniques for identifying specific issues from social media;

- to develop an approach to the automated identification of urban issues;
- to research, select and adapt existing tools to be used for making the developed approach computationally feasible;
- to develop a vocabulary for urban issues;
- to develop a system which implements the proposed approach;
- to adapt a geoparser system to enable geoparsing social media messages for toponyms inside urban areas;
- to analyze social media data from specific urban areas concerning urban issues;
- to analyze the relationship among the location contexts related to a tweet;
- to analyze the relationship between tweet sentiment polarity and urban issues being reported.

1.3 Research Methodology

The research methodology applied in this thesis comprises a set of activities developed in a four-year period. One activity of this research is the investigation of methodologies and techniques for urban issues identification from social media. Such activity was developed during the entire development period of this research, since it is mandatory to keep the related knowledge updated. For this, the main publications related to the field were searched periodically.

Data acquisition is another activity, which involved the investigation of social media networks in order to identify how social media messages could be acquired. Such data were used to evaluate this research. In addition, an investigation searched for urban issues reported on the Web in order to define such domain by means of a vocabulary and concepts.

Another important activity consisted in developing, adapting and improving algorithms that could act on performing the identification of urban issues from social media. Resulting algorithms compose the proposed system. Such an implementation

activity involved not only coding but also performing experiments and conducting case studies to identify problems and overcome the challenges of this research.

Another activity is the evaluation, which consists in performing experiments using collected social media data that could provide evidences to prove the hypotheses concerning this research. The evaluation focuses on measuring the ability of the proposed system in addressing the open problem and comparing the performance with related systems performing the same tasks.

Finally, scientific articles were written and published during the development of this research aiming at sharing the findings with the scientific community. This final manuscript was written and published at the end of this research, answering all the research questions, describing the details of the research steps, discussing the results and related researches to be carried out in the future.

1.4 Contributions

The main contributions of this research include: formally defining the domain of urban issues and discussing the relevance of such domain related to smart cities technologies; presenting a domain ontology designed to enable reasoning in the task of automated identification of urban issues from social media; proposing an automated approach to the identification of urban issues based on domain ontologies; developing a gazetteer enrichment mechanism based on VGI to enable geoparsing in urban areas; and enriching LBSNs in the smart cities domain with linked relevant information from social media streams.

The preliminary results of this thesis were published on both national and international proceedings. The first scientific paper (Oliveira et al., 2014) was published and presented in the 2014 Brazilian Symposium on Geoinformatics (*GeoInfo*), held in Campos do Jordão, São Paulo, with the title "*Automated Production of Volunteered Geographic Information from Social Media*". This publication produced an invitation to submit an extended version (Oliveira et al., 2015a) in the Journal of Information and Data Management, with the title "*Producing Volunteered Geographic Information from Social Media for LBSN Improvement*".

Another paper entitled "*Leveraging VGI for Gazetteer Enrichment: A Case Study for Geoparsing Twitter Messages*" (Oliveira et al., 2015b) was published and presented in the 2015 Wireless Geographical Information Systems International Symposium, held in Grenoble, France. This publication produced an invitation for submitting an extended version in the Journal of Location Based Services, with the title "*Gazetteer Enrichment for addressing Urban Areas: a case study*" (Oliveira et al., 2016).

The paper entitled "*A gold-standard social media corpus for urban issues*" was accepted to be published and presented in the 2017 ACM Symposium on Applied Computing (SAC), to be held in Marrakech, Morocco. The camera-ready version is under production. Another paper entitled "*Building an Ontology for Urban Issue Identification using crowdsourced data*" is being finalized to be submitted to an international journal.

Finally, part of the research for this thesis, the research project entitled "*Generation of Volunteered Geographic Information by inference of Geographical Location extracted from Web documents*", was carried out at University College Dublin (UCD), Republic of Ireland, with co-supervision by Dr. Michela Bertolotto. Such a project was assisted with a one-year scholarship funded by CNPq through the Science Without Borders Programme (*CsF*), process nº 200492/2014-4.

1.5 Document Outline

The remainder of this thesis is structured as follows:

- **Chapter 2 – Background:** discusses the knowledge needed for understanding this work, covering topics such as Volunteered Geographic Information, Geographic Information Retrieval, Location-Based Social Networks and Information Extraction from social media data;

- **Chapter 3 – Related Work:** outlines the state-of-the-art related to complaints identification and urban development projects based on social media data;

- **Chapter 4 – Identifying Urban Issues from Social Media:** presents the proposed approach to the identification of urban issues on social media;

- **Chapter 5 – Evaluation:** describes the case studies carried out in this research and presents the results;

- **Chapter 6 – Conclusion:** presents the conclusion of this thesis and enumerates the further work to be undertaken in the future;

- **Appendix A – Labeling Tweets on Urban Issues and Urban Places:** describes the process performed for labeling the harvested tweets concerning urban issues, geographical locations mentioned and sentiment polarity;

- **Appendix B – Sample Tweet in JSON format:** shows the structure of a tweet by using a true sample harvested from the Web;

- **Appendix C – Social2AGI: A system prototype:** presents a system developed to evaluate the proposed ideas;

- **Appendix D – URL list of FixMyStreet RSS feeds:** lists the URLs used for harvesting the FixMyStreet reports analyzed in this research.

Chapter 2

Background

This chapter presents the main definitions and concepts that are explored during this research. It is structured as follows: section 2.1 presents the location-based social networks and their role in the smart cities domain; section 2.2 discusses the main concepts in Geographic Information Retrieval; section 2.3 presents Volunteered Geographic Information; section 2.4 presents briefly the Twitter social network; section 2.5 presents the main concepts related to information extraction from social media and discusses some approaches from the state-of-the-art; and, finally, section 2.6 presents a summary regarding this chapter.

2.1 Location-Based Social Networks

The Location-Based Social Networks (LBSNs) emerged by combining social networks and Location-Based Services (LBS). An LBSN not only means adding a location to an existing social network (Zheng et al., 2014a). It consists of a social structure made up of individuals connected by the interdependency derived from their location in the physical world as well as their location-tagged media content, such as photos, video, and texts. The interdependency includes not only that two persons co-occur in the same physical location or share similar location histories but also the knowledge (e.g., common interests, behavior, and activities) inferred from an individual's location and location-tagged data.

LBSNs bridge the gap between users' behavior in digital and physical worlds (Cranshaw et al., 2010). People can not only track and share the location-related information but also leverage collaborative social knowledge learned from user-generated and location-related content, such as check-ins, GPS trajectories and geo-tagged photos (Zheng et al., 2014a). LBSNs enable understanding users and location by exploring the relationship between them. For example, LBSNs can be analyzed aiming at estimating user similarity, finding local experts in a region, or inferring location recommendations.

Examples of LBSNs include the Foursquare[3], which gathers consumer experiences in the entertainment domain; Waze[4], which deals with traffic and navigation-related information; and Map My Tracks[5], a GPS activity tracker for fitness purposes. Nowadays, there are some authors that include Facebook and Twitter as LBSN examples since they have introduced location-based features.

2.1.1 LBSN for the smart cities domain

The Location-Based Social Networks for the smart cities domain establish communication between citizens and authorities because the shared content is related to the city's infrastructure. While the citizens act complaining through urban issues reports, the authorities become aware and then can act in order to fix them and also provide replies. Examples of LBSNs in the smart cities domain include FixMyStreet (Walravens, 2013), Crowd4City (Falcão, 2013), ImproveMyCity (Tsampoulatidis et al., 2013), among others.

FixMyStreet (Walravens, 2013) is an open source Web platform that enables users to complain about urban issues. It is a part of the mySociety project of UK Citizens Online Democracy that has been used by many urban areas in countries around the world. The United Kingdom instance[6] is in use by hundreds of councils, metropolitan or non-metropolitan districts that granted city status. The instance running in the Republic of Ireland[7] is used by 34 councils.

A screenshot of the FixMyStreet instance running in Dublin City, Ireland, is shown in Figure 2.1.

[3] http://www.foursquare.com/
[4] http://www.waze.com/
[5] http://www.mapmytracks.com/
[6] http://www.fixmystreet.com/
[7] http://www.fixmystreet.ie/

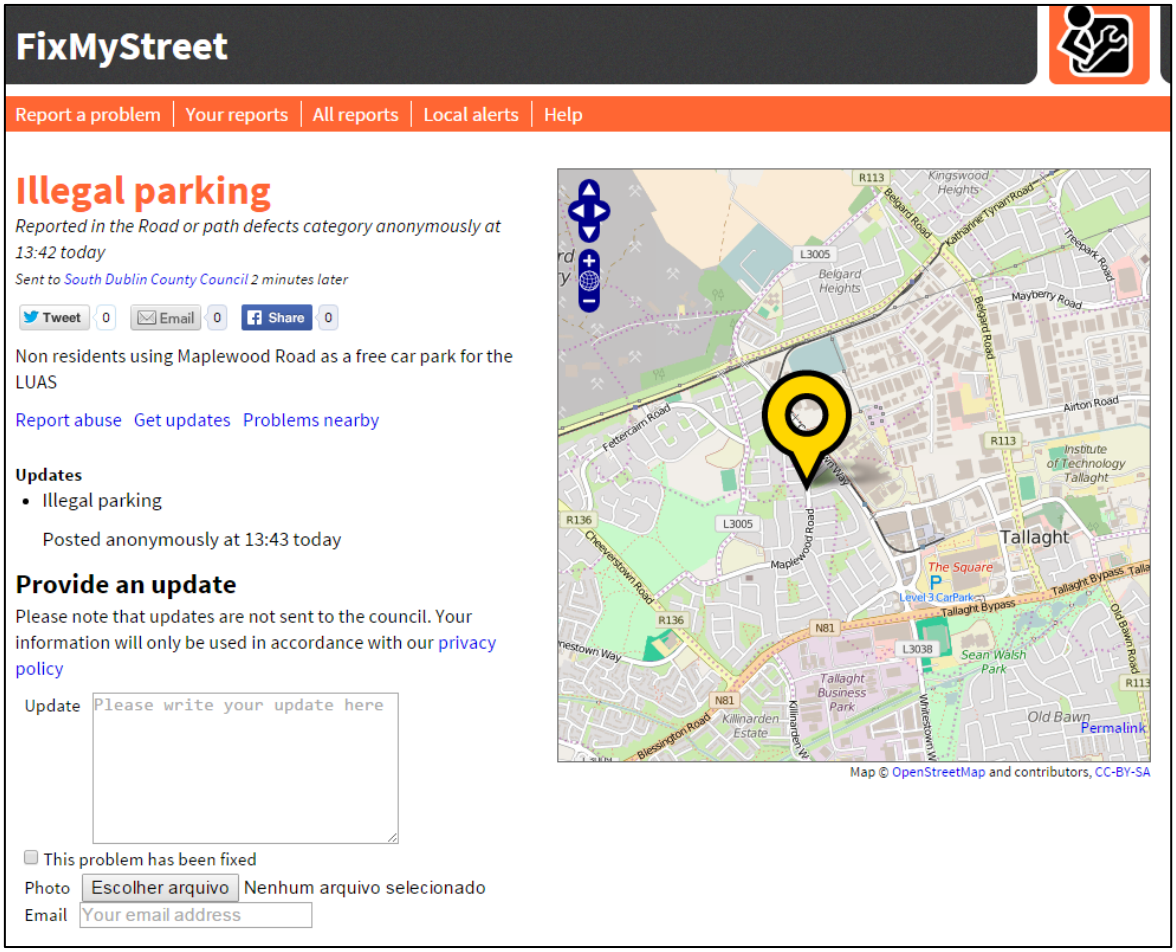


Figure 2.1: An example of issue reported in the FixMyStreet running in Ireland

In the example shown in Figure 2.1, an anonymous user is reporting an illegal parking in a specific location of Dublin City. FixMyStreet shows the geolocated content in a map and offers an option to other users share updates regarding the issue, including photos of the problem and reports that the problem was fixed. FixMyStreet also shows a status regarding the urban issue that is automatically sent to the local council, which is in charge of the area related to the issue.

ImproveMyCity[8] (Tsampoulatidis et al., 2013) is another open source platform that enables citizens to directly report to their public administration local issues about their neighborhood. Examples of such reported issues are faulty street lights, broken tiles on sidewalks and illegal advertising boards. Similarly to FixMyStreet, ImproveMyCity automatically transmits the shared information to the appropriate office in public administration. Figure 2.2 presents a screenshot of the ImproveMyCity live demo instance.

[8] http://www.improve-my-city.com/

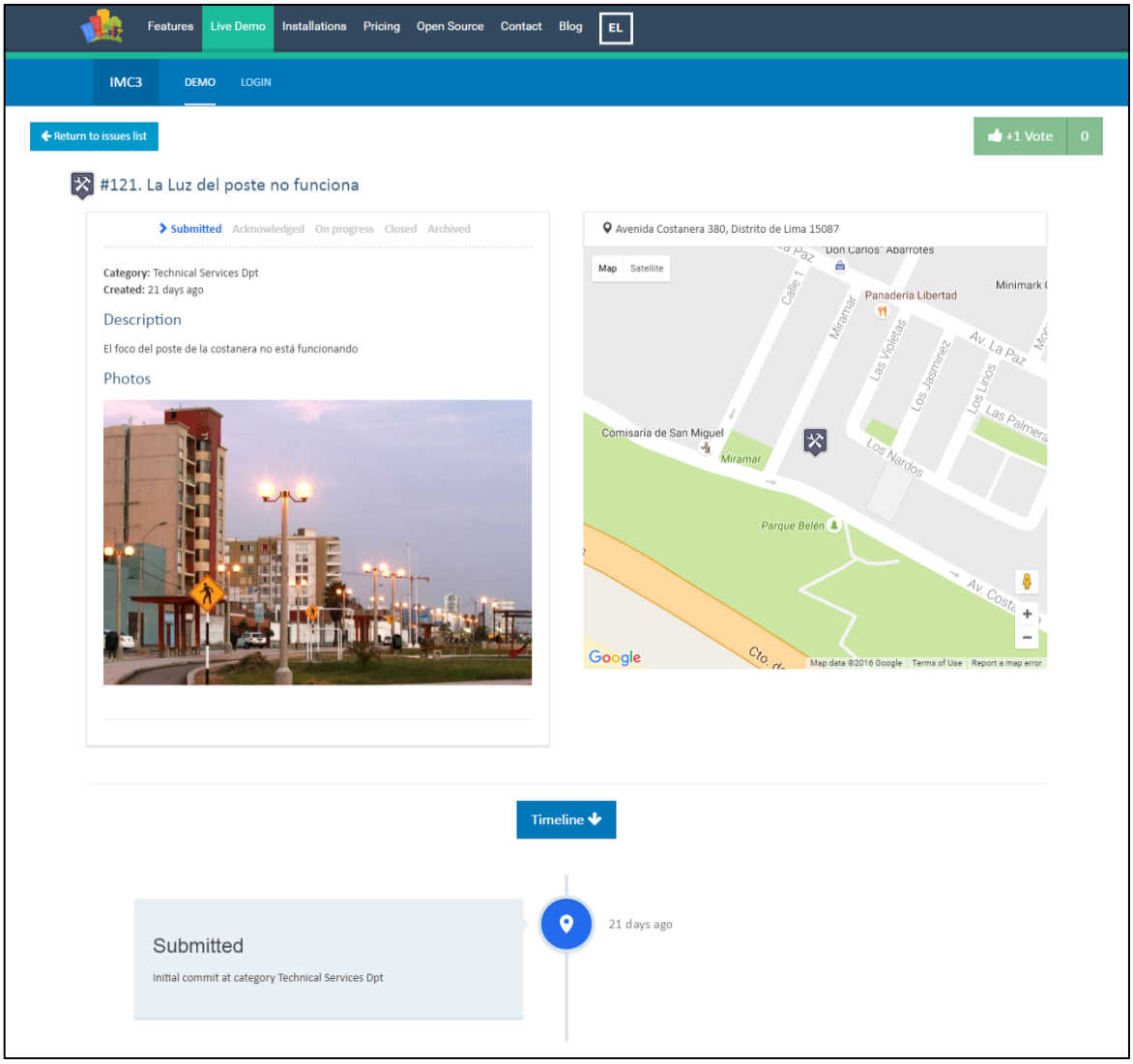


Figure 2.2: An example of issue reported in the ImproveMyCity live demo

The example shown in Figure 2.2 consists of an urban issue report in Spanish about faulty street lights. ImproveMyCity provides a timeline feature that shows the historic of the urban issue as its status change along the time. An urban issue report may present one of the following statuses: submitted; acknowledged; on progress; closed; and archived.

Crowd4City (Falcão, 2013) is a Brazilian LBSN designed for citizens to report urban issues and being connected with public administration agencies. The Crowd4City LBSN is a Web application that works similarly to FixMyStreet and ImproveMyCity, but presents an interesting feature that enables users to report their issues using not only a point but also lines and polygons for the geographical reference. A screenshot of the

Crowd4City instance running in the city of Campina Grande, in Brazil, is shown in Figure 2.3.

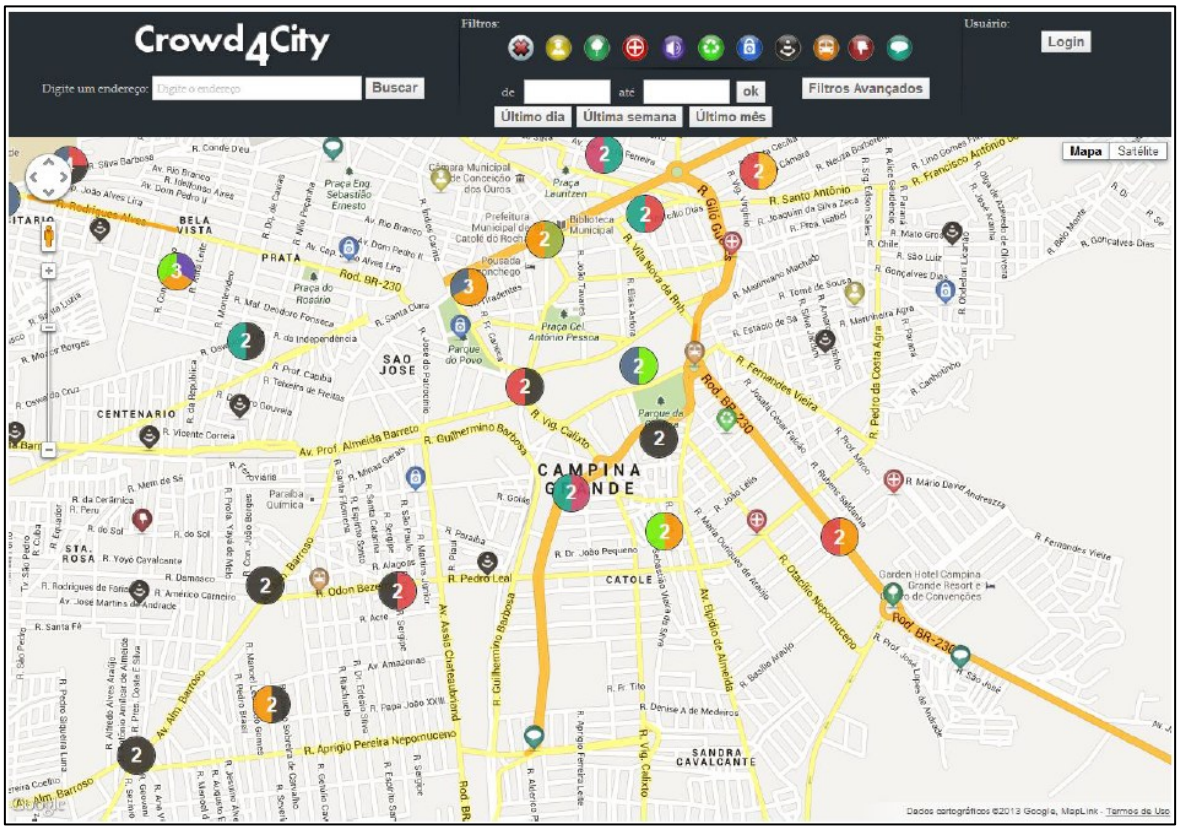


Figure 2.3: A screenshot of the Crowd4City LBSN (Falcão, 2013)

In the Crowd4City LBSN map, the colored clusters illustrate the urban issue types in a region of the city according to the amount of occurrences. The greater the number of the reports for a given issue type is, the larger the slice size in the cluster for the representative of such issue type. Thus, it is easy to identify regions of the city where the transportation is the main issue type, for instance, just looking for the orange color in the clusters.

Colab[9] (Garcia et al., 2016) is another Brazilian LBSN designed for monitoring urban issues reported by the citizens. Similar to the Crowd4City, the Colab is connected with the city halls in order to make them aware regarding the upcoming urban issues that require solutions. The city halls can also return answers to the citizens regarding the issues reported. Currently, around 30 Brazilian cities are using the Colab.

[9] http://www.colab.re/

Figure 2.4 shows the Colab instance running in the city of Natal, in Brazil, with an urban issue report about rubbish in the streets. The citizens from the cities served by the Colab can report the urban issues through a Web application or an app for mobile devices.

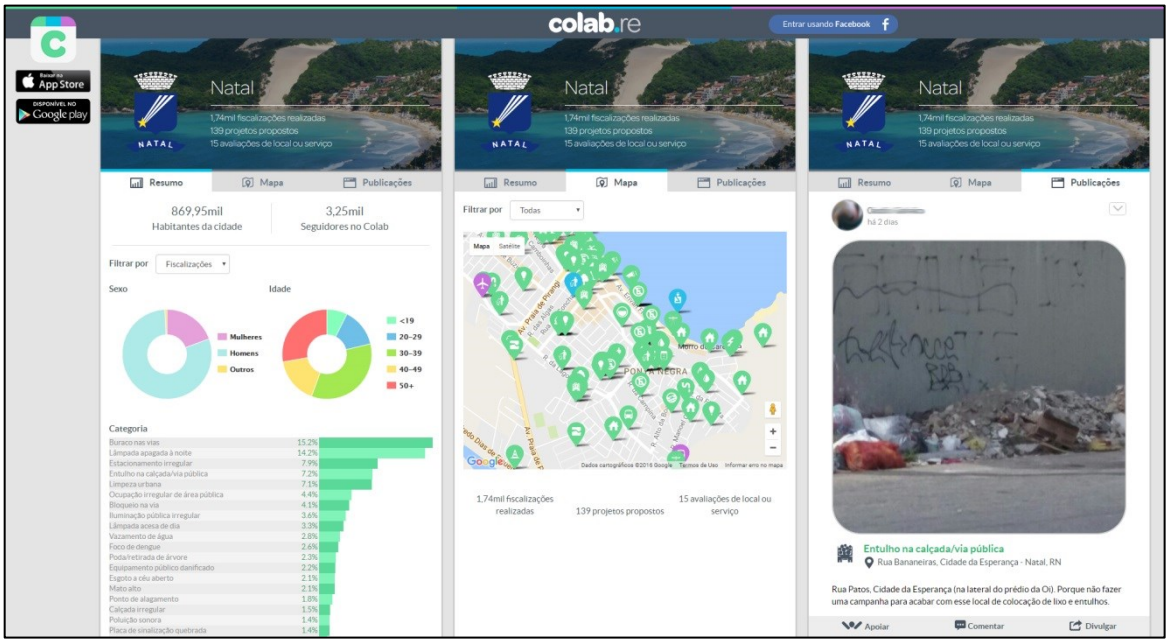


Figure 2.4: A screenshot of the Colab LBSN Web application

Other Brazilian LBSN in the domain of smart cities is the *ReclameAquiCidades*[10] (that means "*Complain here about cities*" in Portuguese). The *ReclameAquiCidades* was created as a specialization of the *ReclameAqui*, a Brazilian website for customer complaints, in order to gather complaints about general public services and urban issues in Brazilian cities. Currently, around 120 Brazilian cities are using such an LBSN. Unlike the others LBSN shown, the *ReclameAquiCidades* does not have an interactive map to show the city overview regarding the reported issues.

Figure 2.5 presents a screenshot from the *ReclameAquiCidades* showing the summary of 5,868 urban issues reported in the city of São Paulo[11], in Brazil, per issue type such as "*Trânsito*" (Traffic), "*Educação*" (Education), "*Saúde*" (Health), "*Buracos*" (Potholes), among others. As it can be seen, only around 10% of the urban issues reported were answered by the city authorities. It suggests the connection between the online citizens and city authorities is not working as expected.

[10] http://cidadao.reclameaqui.com.br/
[11] http://cidadao.reclameaqui.com.br/indices/2/prefeitura-sao-paulo/

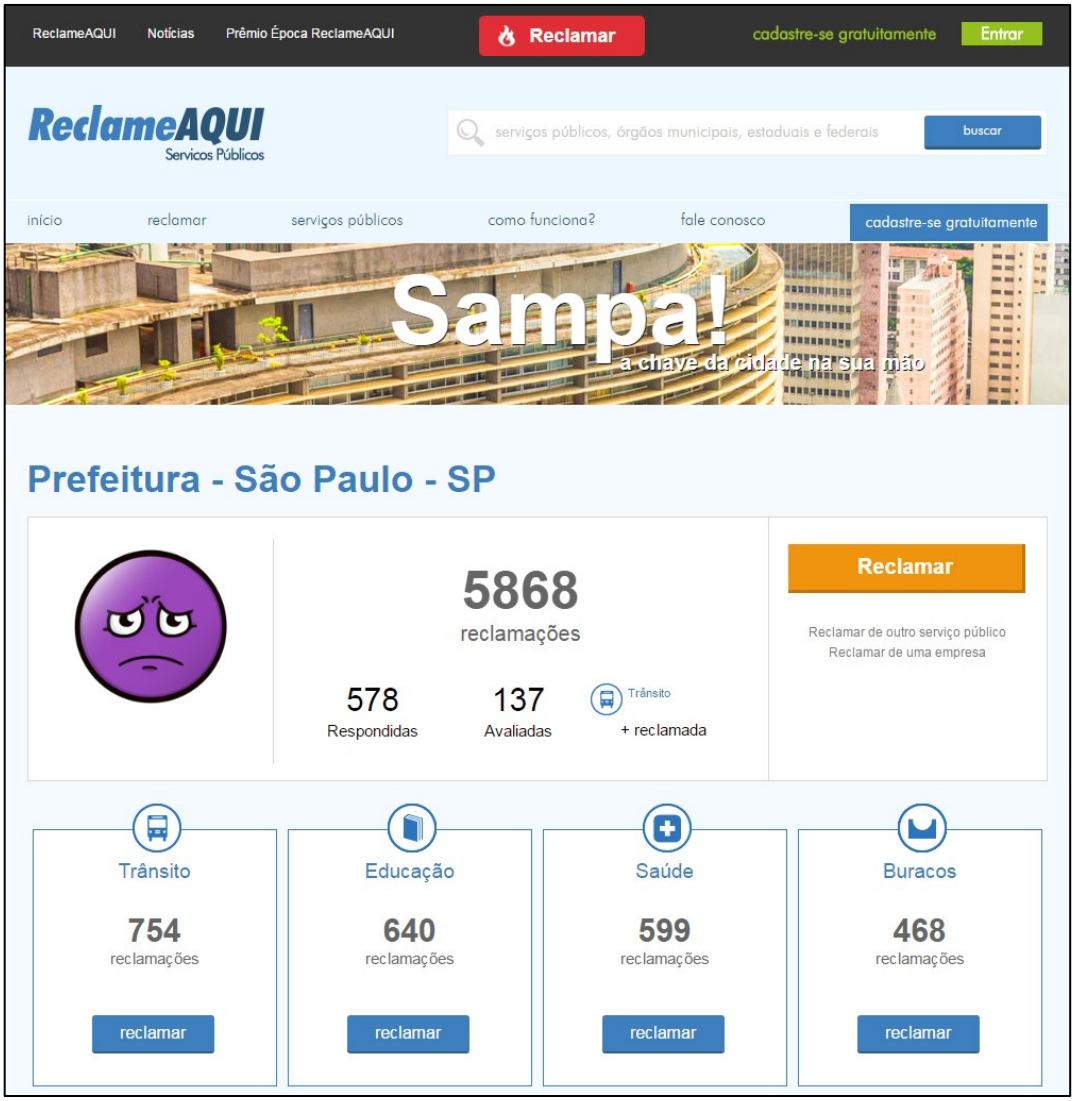


Figure 2.5: A screenshot of *ReclameAquiCidades* in São Paulo, Brazil

BuitenBeter[12] (Sidawy, 2010) is a Dutch LBSN that provides a direct and relevant communication channel between citizens and local government about issues regarding the public space. Users are enabled to report urban issues through smartphones such as potholes, stray garbage and broken street lamps. The main difference from the others LBSNs shown is that BuitenBeter is a mobile device app, while the others are Web applications.

The interfaces for two BuitenBeter screens are illustrated in Figure 2.6. In order to report an urban issue using BuitenBeter, the user can take a picture, confirm the geographical location on a map, select a category for the issue and write a description.

[12] http://www.buitenbeter.nl/

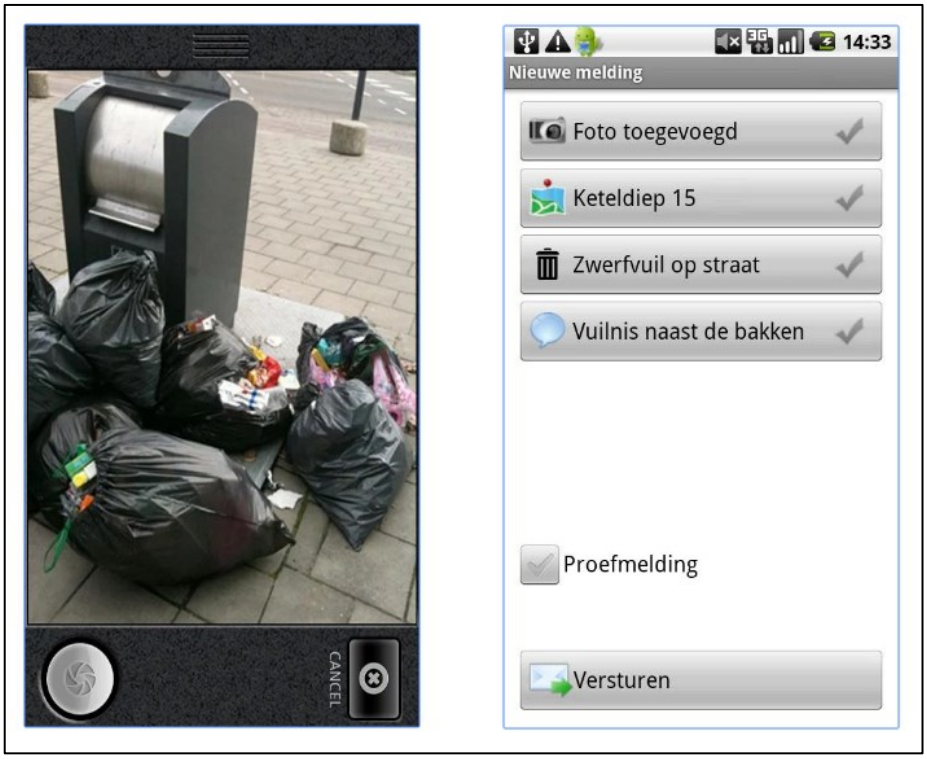


Figure 2.6: BuitenBeter screens (from Google Play[13])

Focusing on the subdomain of public security in urban areas, the LBSN OndeFuiRoubado[14] (that means "*Where I was robbed*" in Portuguese) has been used by around 800 Brazilian cities. Different from the LBSNs previously presented, OndeFuiRoubado users are in charge of complaining only about security issues such as thefts, robberies and break-ins. Those issues are sent to the local Police and other government security agencies. Figure 2.7 shows a screenshot of the *OndeFuiRoubado* Web application with information from the city of Campina Grande, in Brazil.

Other LBSN example in the subdomain of public security in urban areas is the WikiCrimes (Furtado et al., 2012), another Brazilian LBSN which contains over 280,000 reported crimes. Figure 2.8 shows a screenshot of the WikiCrimes LBSN running in the city of Fortaleza, in Brazil.

We could notice that many LBSNs around the world share similar purposes. Some cities may present more than one LBSN that overlap or complement each other. Overlapping LBSNs in a unique city is not good because it is a challenge for the local

[13] https://play.google.com/store/apps/details?id=com.yucat.buitenbeter
[14] http://ondefuiroubado.com.br/

authorities, which needs to check each one and remove duplicate issues. Therefore, solutions such as the Dublin Dashboard have emerged aiming at integrating different LBSNs in the smart cities domain.

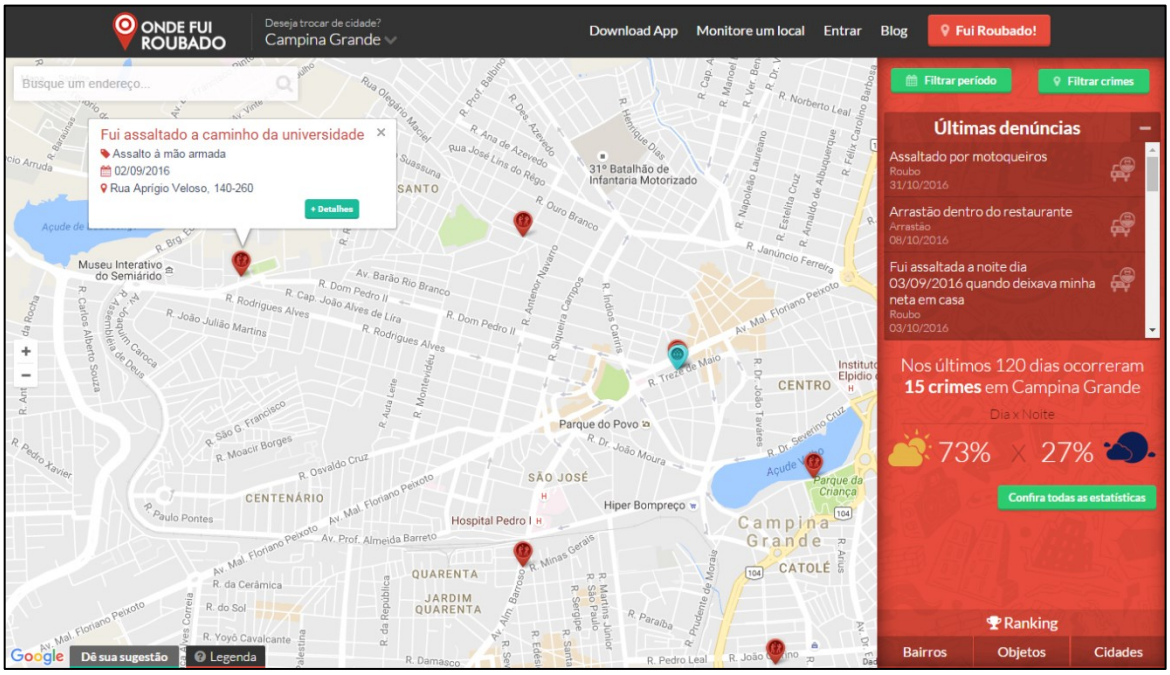


Figure 2.7: A screenshot of the *OndeFuiRoubado* LBSN

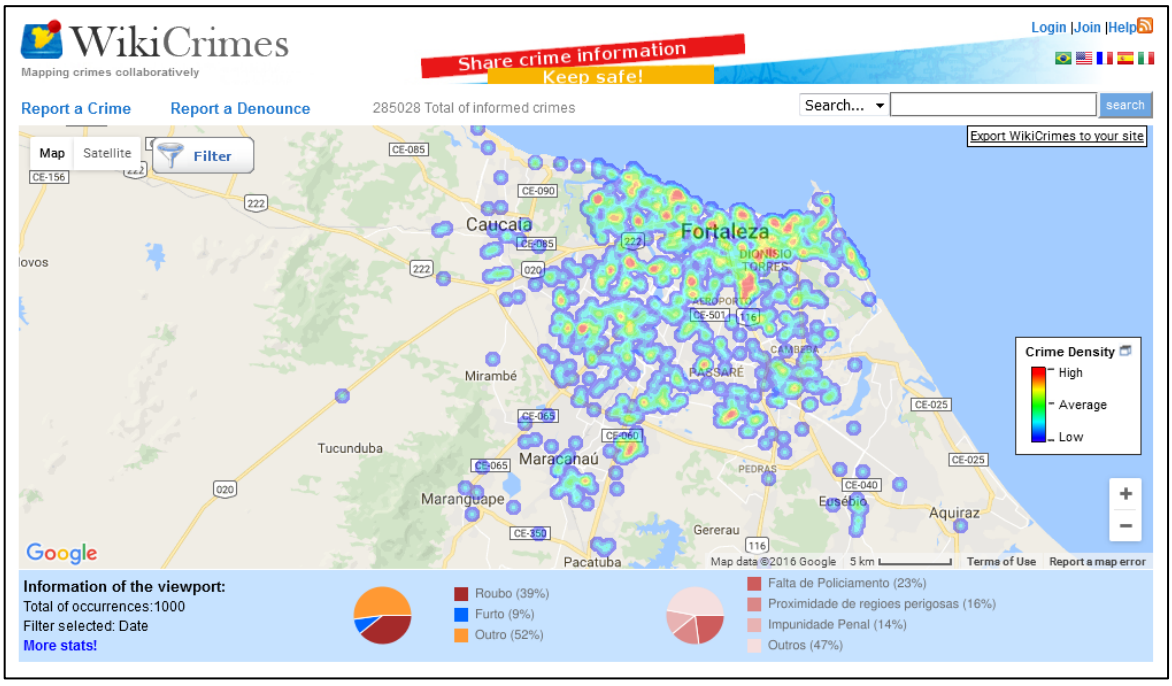


Figure 2.8: A screenshot of the WikiCrimes LBSN

The Dublin Dashboard[15] is an analytical dashboard that provides city managers and citizens with up-to-date detailed information about different aspects of urban areas (Kitchin et al., 2015). This dashboard gathers several different LBSNs developed to deal with specific issues in the Dublin City, in Ireland, such as planning, housing, transportation, security and urban issues. Figure 2.9 illustrates the main screen of the Dublin Dashboard.

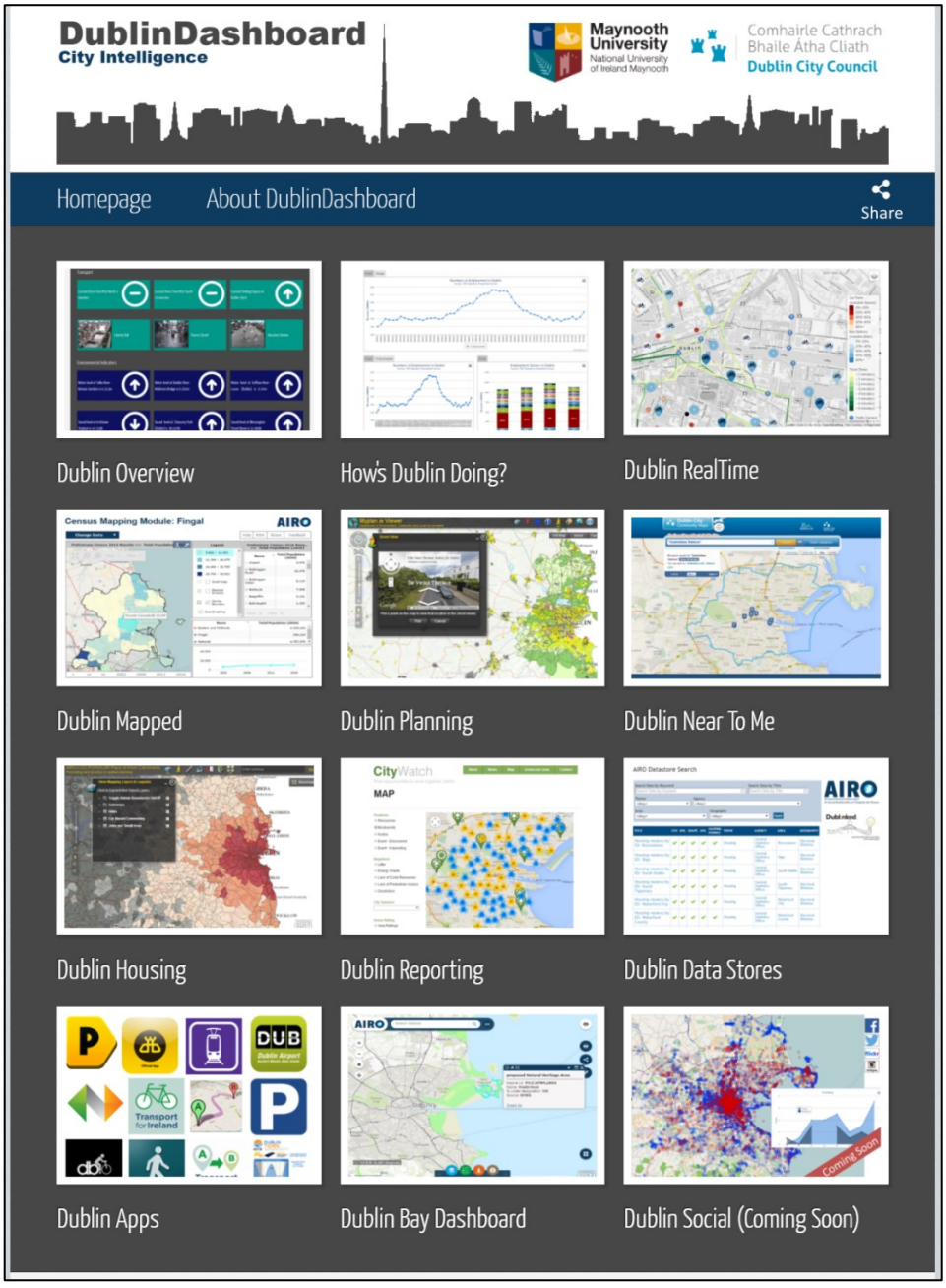


Figure 2.9: The main screen of the Dublin Dashboard

[15] http://www.dublindashboard.ie/

Besides FixMyStreet, Dublin City has three other LBSNs which deal with urban issues: CityWatch[16], FixYourStreet[17] and FixMyArea[18]. Different platforms sharing the same purpose being used in the same urban area will both confuse the citizens and make it difficult for the councils to collect all the information. Such platforms would be more useful if they were integrated or even merged.

2.2 Geographic Information Retrieval

Geographic Information Retrieval (GIR) is an ongoing research field that is part of the broader research stream of Information Retrieval (IR). GIR comprises many proposals addressing geographical information extraction on semistructured documents using Natural Language Processing (NLP) and geoprocessing techniques. GIR algorithms are designed to process text from Web resources (e.g., Webpages, blogs and social networks) and then assign specific geographic locations to them (Purves and Jones, 2011).

Within this context, gazetteers are known as huge geographical knowledge databases and have been fundamental for GIR tasks in order to connect place names (also known as toponyms) to geographical features or footprints (Keβler et al., 2009). The wide variety of such gazetteers is enriched by geography experts endowed with skilled knowledge in order to provide geographical data of acceptable quality. Although desirable, this way of spatial data production is costly, requires time and there is a lack of free services availability. GeoNames[19] is one of the most known open gazetteers and its database covers all countries and contains over 10 million place names and over 9 million unique features[20].

The problem with current open gazetteers is that they do not cover geographical features at a high Geographic Level of Detail (GLoD), e.g. Points-Of-Interest (POIs), Streets and Districts. Considering that gazetteers are mainly used by geoparser systems, which work aiming at identifying geographical locations in machine-readable texts, those systems are only able to resolve toponyms relying on the GLoD provided by the gazetteers.

[16] http://www.citywatch.ie/
[17] http://www.fixyourstreet.ie/
[18] http://fixmyarea.com/
[19] http://www.geonames.org/
[20] Information updated in November 2016.

A geoparser system (or just geoparser) generally performs the identification of geographical locations in two stages: toponym recognition (geoparsing) and toponym resolution (georeferencing). Geoparsing is in charge of identifying candidate terms that might refer to a geographical location. Georeferencing is in charge of assigning real-world coordinates to candidate terms. Those candidate terms from a text with coordinates assigned to them are toponyms resolved from such text. Geoparser systems rely on one or more gazetteers, which are fundamental on both stages. It is common to find in the literature the term geoparsing meaning the entire GIR processing. Figure 2.10 illustrates a geoparser system processing a Twitter message.

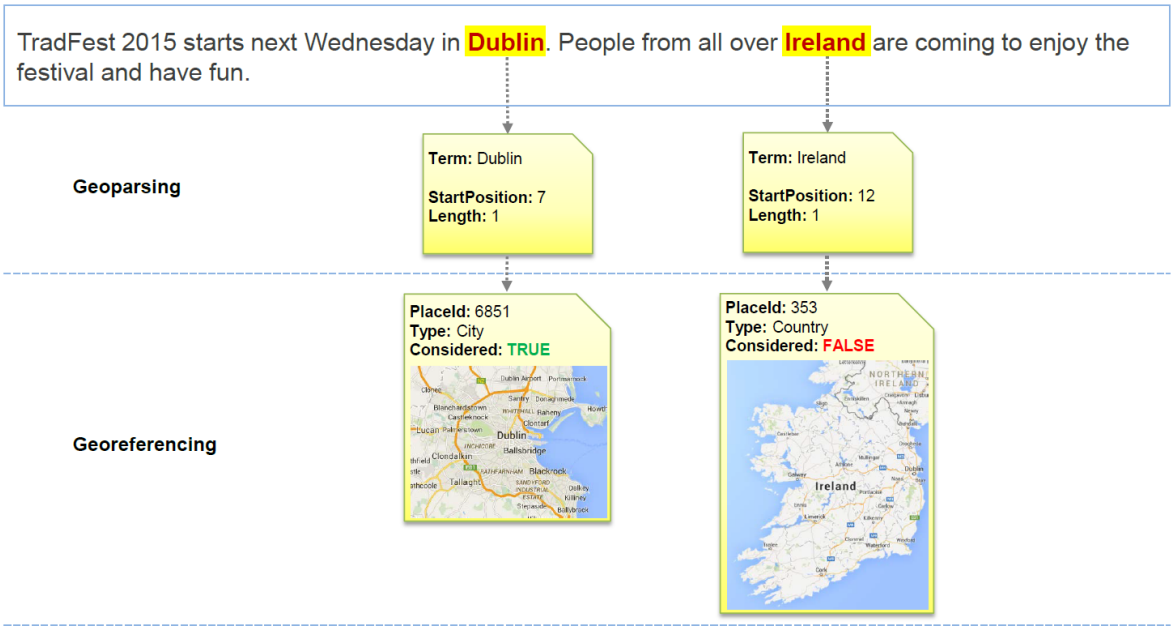


Figure 2.10: Illustration of geoparsing and georeferencing stages applied to a Twitter message

In Figure 2.10, two candidate terms from the processed tweet are identified by the geoparser during the geoparsing stage. However, only one of them (Dublin) is a toponym resolved after the georeferencing stage has finished. Although the other candidate term (Ireland) also refers to a geographical location (a country), the adopted georeferencing strategy discarded such toponym in order to keep only the most specific one due to Dublin City is located inside Ireland.

In the following, the GeoSEn system, a system which performs GIR and is much explored along this research, is presented.

2.2.1 The GeoSEn system

Campelo and Baptista (2009) proposed a model for extraction of geographic knowledge from Web documents. They developed GeoSEn, a search engine with geographic focus, which enables the geographic indexing of documents extracted from the Web. The architecture of the GeoSEn system was developed as an extension of the Apache Nutch Framework[21], by adding the ability to manipulate and retrieve geographic information. Since Nutch is based on a plugin-oriented architecture, some plugins were implemented and they are the essence of the GeoSEn system. This section focuses on the GeoSEn Parser, one of these plugins, since it is related to this research.

The GeoSEn Parser is a geoparser responsible for the detection of geographic terms from texts written in Portuguese. It enables to infer toponyms eventually cited in a text following the Brazilian political-division hierarchy, which goes from the least precise levels (countries) to the most precise ones (cities). The GeoTree (Campelo, 2008) is a tree-based data structure that establishes the hierarchical relationship between toponyms stored into the GeoSEn gazetteer. There are six levels of toponyms in the GeoTree: Country, Region, State, Mesoregion, Microregion and City. An example of a GeoTree instance is illustrated in Figure 2.11.

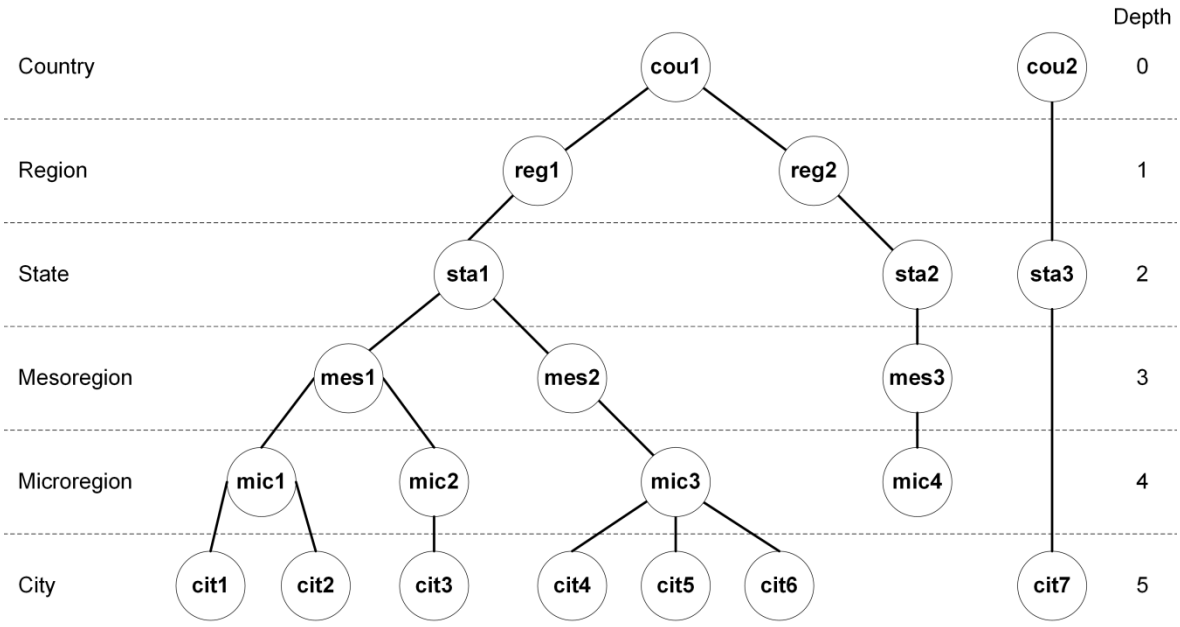


Figure 2.11: An example of a GeoTree instance (adapted from Campelo, 2008)

[21] http://lucene.apache.org/nutch

The process of detecting geographic references is based on a set of heuristics. For example, the geoparser considers information such as the position of a term in the text and its length (i.e. the number of words that form a term). Such information about position of the terms can be useful to correlate spatial terms which are close in the messages.

The GeoSEn system includes a gazetteer which is composed of toponyms from Brazil, according to the GeoTree structure. Each toponym is stored by `loc:id`, the unique identification; `loc:type`, the type of the toponym, which means the respective level in the GeoTree; `loc:name`, the full name; `loc:ancestor`, the identification of the toponym that is a direct ancestor; and `loc:geometry`, the embedded geometry using the projection EPSG:4326, which is related to the WGS84 geodetic system adopted by Google Maps and GeoNames.

The GeoSEn system uses a metric called Confidence Rate (CR). The CR is a measure that represents the probability of a toponym resolved candidate to be a valid place. The CR value varies between 0 and 1, inclusive. In the GeoSEn system, each toponym level in the GeoTree has a specific CR calculation. A specific CR for each toponym type is necessary as there are several different ways that influence people mentioning a toponym depending on such type.

The CR values are calculated based on Confidence Factors (CF). A CF is a measure associated with an analyzed feature of a toponym during the parsing process. The GeoSEn system uses four different CFs:

- CF_{ST} : related to the occurrence of a special term[22] before or after the term that results in a toponym-resolved candidate;
- CF_{FMT} : related to the spelling correctness of a toponym-resolved candidate when it is compared to terms from the source text;
- CF_{CROSS} : related to other toponym-resolved candidates;
- CF_{TS} : related to textual searches performed by the GeoSEn search engine.

[22] Special terms are words (or terms) that frequently appear before or after a toponym (e.g. “in”, “at”, “near”) and can be used in order to help the toponym recognition task in the geoparsing process.

2.3 Volunteered Geographic Information

Volunteered Geographic Information (VGI) has emerged in the last years as an alternative and powerful spatial data source on the Web. VGI consists of geographic data provided voluntarily by individuals who act as sensors (Goodchild, 2007). Since the main feature of the crowdsourcing phenomenon is the dissemination of data produced by people spread around the world, VGI emerges as a specific kind of crowdsourced data. Most of these volunteers are ordinary people interested in sharing their footprints, viewpoints and knowledge about geographic locations.

In VGI, the information shared by volunteers is linked to a specific geographic region. While such information is usually related to elements of traditional cartography, VGI also may include subjective, emotional, or other non-cartographic information (Parker, 2014).

Research on VGI has prevailed in several parts of the world. Besides Computer Science, many correlated disciplines, such as Geography (Goodchild, 2007), Geographic Information Science (Jackson et al., 2010) and Human Factors (Parker et al., 2011) have investigated issues concerning this kind of volunteered information.

One of the most representative VGI project is the OpenStreetMap[23] (OSM) (Haklay and Weber, 2008). The OSM database consists of a significant collection of VGI based on the Wikipedia collaborative model (Mooney and Corcoran, 2012). OSM contains approximately 2.8 billion spatial features or 500 GB of XML data[24] available to be freely downloaded and used; and approximately 1.8 million registered users, nevertheless only 25% are considered active contributors[25]. Thus, OSM is the current reference for VGI worldwide.

The OSM project has received many contributions from the community. Haklay (2010), for instance, has focused on assessing VGI quality and how VGI from OSM can be reliable and usable. Ballatore and Bertolotto (2011) focused on semantic relationships within OSM data. They highlight how OSM is spatially rich but semantically poor and investigate ways of linking OSM to other distributed repositories.

[23] http://www.openstreetmap.org/
[24] http://wiki.openstreetmap.org/wiki/Planet.osm
[25] http://wiki.openstreetmap.org/wiki/Stats

Besides the OSM project, several works have revealed VGI as a promising research field. Horita et al. (2013) made a thorough literature review on VGI with the objective of verifying its applicability for aiding in disaster management. Such study shows that the VGI has been more frequently used in fires and floods. Havlik et al. (2013) discussed VGI mobile applications concerning several aspects, such as functionalities and user experience. Ballatore et al. (2013) explored the semantic side of VGI and presented a technique for computing the semantic similarity of geographic terms in VGI based on their lexical definitions and using WordNet[26]. The authors based themselves on the intuition that similar terms tend to be recursively defined by similar terms. Some other examples of VGI projects include WikiMapia[27], Google Map Maker[28], and Waze.

2.3.1 Ambient Geographic Information

Geo-referenced data produced within services such as Trip Advisor[29], Flickr[30], Twitter and Panoramio[31] can also be considered VGI. However, according to Stefanidis et al. (2013), the information disseminated through social media is a deviation from VGI since it is not geographic information per se. They agree that geographical information may be present in social media data in several manners. Nevertheless, social media data tends to be Ambient Geospatial Information (AGI) instead of VGI, since it goes beyond geography and represent human activities. While VGI is provided by active volunteers, the concept of AGI relates to ambient sensors which act passively (Crooks et al., 2015).

2.4 The Twitter

A challenging aspect of processing social media, particularly microblogs such as Twitter, is the short length of the messages. Twitter is one of many social media networks running worldwide. Twitter is also called microblog due to the limitation of 140 characters (about 24 words) in a message, called tweet and sometimes mentioned as Twitter post. Currently, more than 500 million of tweets are posted in a unique day[32].

[26] http://wordnet.princeton.edu/
[27] http://wikimapia.org/
[28] https://www.google.com/mapmaker
[29] https://www.tripadvisor.com.br/
[30] http://www.flickr.com/
[31] http://www.panoramio.com/
[32] http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html

Twitter was chosen for this research project since it allows the freedom of users writing about any topic and there is a quantity of data freely available to researchers (Gelernter and Mushegian, 2011). A tweet is time-stamped automatically with Greenwich Mean Time (GMT). Other factors have contributed to the wide usage of Twitter on researches related to social media data, such as robustness, to be up-to-the minute and handle high volume data. Moreover, Twitter provides an Application Programming Interface (API) – the Twitter API[33] – that enables access to public tweets with several constraints regarding the volume and the age of the accessed tweets.

Although such Twitter data access constraints may represent a trouble for research purposes, Twitter is currently the most used social media network by the community (Alves et al., 2015; Anantharam et al., 2015; Wakamiya et al., 2015; Xia et al., 2015; Eshleman and Yang, 2014). The widespread usage of Twitter in researches worldwide can be explained by the challenges faced on trying to access data from other social media such as Facebook. However, such challenges may be further minimized as social media networks are continuously improving their APIs.

A concept related to tweets is the hashtag, a type of label that makes it easier for users to find messages with a specific topic. Hashtag became a practice of writing style on social media. Hashtags are created by placing the hash character # in front of a word or unspaced phrase. A detailed example of a tweet in JSON (JavaScript Object Notation) format provided by the Twitter API is shown in details in Appendix B.

2.5 Information Extraction from Social Media

Social media networks have increased in popularity during the last few years. Since there are millions of people talking about everything in an unstructured way, social media have become a rich source of data for research around the world (Chan et al., 2014). Each social media user can be seen as an agent or sensor that shares information continuously (Gelernter and Mushegian, 2011). In addition to being huge, the flow of information is informal, noisy and about different domains, which makes evident the need to be mined to reveal relevant information for specific purposes.

[33] http://dev.twitter.com/overview/api

Such context involves an Information Extraction problem. Information Extraction (IE) is the problem of summarizing the essential details particular to a given document (Freitag, 2000). In social media, a post is the message shared by the users. Such a post can be analogously seen as a document for IE purposes. IE tasks extract structured information from unstructured or semi-structured machine-readable documents.

Usually, IE concerns processing human language texts by means of NLP. Some known IE tasks are Named Entity Recognition (NER), Named Entity Disambiguation (NED), Named Entity Linking (NEL) and vocabulary analysis. Such IE tasks can be used separately or combined in IE approaches. Two standard IE approaches involves machine learning classifiers (e.g. Naïve Bayes, Support Vector Machines) and sequence models (e.g. Conditional Markov Model, Conditional Random Fields).

Extensive research into classifiers for short texts such as tweets has concluded that classification and information extraction from social media is significantly harder than for longer documents (Chenliang et al., 2012). Whilst traditional text mining techniques are based on more structured and less noisy data than social media, novel or adapted techniques needed to be developed in order to overcome the challenges posed by such kind of information.

Many proposals have arisen to deal with information extraction from social media in three contexts: thematic (also seen as topical), geographical and temporal. Stefanidis et al. (2013), for example, discussed the possibilities of analyzing information from social media considering space, time, and topic. The topical information relies on a set of keywords, whilst the spatial information looks into the metadata to identify the geocoding information and, finally, the temporal information relies on the timestamp when the tweet was published.

Information Extraction from social media has been applied in different purposes such as event detection, sentiment analysis, user analysis, among others. Event detection focuses on discovering social media streams which may be talking about specific facts in the space-time, such as forthcoming events or a natural hazard impact. Sentiment analysis focuses on the mood of social media, developing techniques to classify social media messages into Positive, Negative or Neutral (Alves et al., 2015; Wakamiya et al., 2015;

Eshleman and Yang, 2014). User analysis focuses on both message and user metadata in order to discover mobility and density patterns (Mahmud et al., 2014).

This section keeps the focus on event detection research in order to discuss techniques that might be used for urban issues identification from social media. Before exploring the state of the art in event detection from social media, the entity recognition needs to be introduced, since it is a primary task in information extraction.

2.5.1 Entity Recognition and Disambiguation

Entity recognition (or entity extraction), widely known as Named Entity Recognition (NER), is a IE subtask that seeks to locate and classify entities mentioned in a text according to previously-defined classes such as Person, Organization and Location. While NER focuses on identifying the mention of an entity in a text, the Named Entity Disambiguation (NED) is another IE subtask that focuses on identifying which specific entity it is.

For example, the term "São Paulo" may refer to a Brazilian city or state (Location), to a football team (Organization) or even to a holy person (Person). Thus, while NED focuses on identifying the correct entity according to the context in the text, NER does not take such particularity into account and may identify one of the three possible entities randomly. Some articles from the literature may also refer to NER and NED as Named Entity Recognition and Disambiguation (NERD), a combination of both subtasks.

While NER is designed for many purposes, event detection can be seen as a specialization of entity recognition, since it explores and combines entity recognition techniques for specific domains and scenarios. There are a number of systems which perform NER/NED from a given dataset for different purposes such as the online services AlchemyAPI[34], OpenCalais[35], Zemanta[36] and the demo provided by the Stanford Group[37]. The main issue regarding such services is the imposed restriction concerning the amount of data to be processed (Feyisetan et al., 2014).

[34] http://alchemyapi.com/
[35] http://opencalais.com/
[36] http://zemanta.com/
[37] http://nlp.stanford.edu:8080/ner/process

According to Lingad et al. (2013), other NER tools such as Stanford NER[38], OpenNLP[39], Twitter NLP (Ritter et al., 2011) and Yahoo! PlaceMaker[40] can be applied for extracting locations from tweets. However, experiment results have shown that the Twitter NLP surprisingly performed worse than Stanford NER (Lingad et al., 2013) even though Twitter NLP was developed to handle tweets as opposed to Stanford NER. The importance on using appropriate training data on those NER/NED tools could be noticed. Regarding applying NER on tweets, hashtags need to be considered since it also may include location references. Moreover, the effectiveness of NER/NED tools is affected by the quantity of training data. This fact is a challenging issue because it is difficult to provide enough manually annotated tweets for training.

The expensive human effort required from TwitterNLP due to its supervised feature was reported by Li and Sun (2014). They developed a time-aware POI tagger based on TwitterNLP and the Conditional Random Field (CRF) Model with the BILOU schema. Once a POI had been mentioned in a tweet, the Li and Sun (2014) tried to predict whether such tweet refers to a past experience, a current check-in or a future plan about going to such place, using the Foursquare POI's database as a gazetteer. The main challenge faced by Li and Sun (2014) was due to people often mentioning POIs with abbreviations or partial names, assuming awareness from the audience.

CRF is a framework for building probabilistic models to segment and label sequence data (Lafferty et al., 2001). CRF is generally applied in machine learning or pattern recognition for structured prediction. For example, in a social media post, a CRF model may apply the label "NOUN" to a word based on the neighboring words already labeled as "ARTICLE" (the previous one) and "VERB" (the subsequent one). Such task of grammatical tagging or word-category disambiguation is known as part-of-speech tagging (POS tagging) and is performed by POS-Taggers. The BILOU schema is an approach which helps labeling compound words. The BILOU acronym comes from **B**eginning, **I**nside and **L**ast tokens of multi-token text segments, **U**nit-length text segment and **O**utside of a text segment. The BILOU schema significantly outperforms the BIO schema (**B**eginning, the **I**nside and **O**utside of text segments) (Ratinov and Roth, 2009).

[38] http://nlp.stanford.edu/software/CRF-NER.shtml
[39] http://opennlp.apache.org/
[40] http://developer.yahoo.com/geo/placemaker/

2.5.2 Event Detection

An event can be seen as something that happens in a region during a period of time. Events might have actively participating agents, passive factors, products, and a location in space/time (Raimond and Abdallah, 2007). Research on event detection from social media generally targets events such as earthquakes, hurricanes, storms, musical gigs, sports competitions, accidents, traffic jams and political campaigns. Such events have several properties (Sakaki et al., 2010): i) they may be of large scale, ii) they may particularly influence people's daily life, and iii) they may have both spatial and temporal regions.

A survey of techniques for event detection from Twitter feeds is presented by Atefeh and Khreich (2013). It discusses the challenge of performing event detection in tweets due to the diversity of users, the usage of informal, irregular and abbreviated words, and the large number of spelling and grammatical errors, improper sentences, mixed languages, rumors and meaningless messages. These characteristics make traditional text mining techniques inappropriate for this data.

Event detection techniques can be classified according to the event type (specified or not), detection method (supervised or not), detection task (retrospective or new events) as well as the target application (Atefeh and Khreich, 2013). Concerning unspecified event detection, most current techniques applied to Twitter streams rely on clustering approaches. This trend can be explained by the fact that most clustering algorithms do not require prior knowledge. On the other hand, most techniques for specified event detection rely on models of previous and static knowledge about an event, although some incremental learning approaches have been proposed as well. Basically, a classifier is typically trained off-line on a relatively small tweet dataset manually labeled and combined with a clustering approach.

Atefeh and Khreich (2013) pointed out the considerable effort needed to achieve efficient and reliable event detection systems, including: designing better feature extraction and query generation techniques; developing accurate filtering and detection algorithms; improvement of techniques to combine and analyze information from multiple sources and multiple languages; and enhancing summarization and visualization approaches. They also

discussed cost issues suggesting the adoption of preprocessing filters in order to remove non-relevant messages and reduce the amount of data to be processed.

A framework for automatically collecting and filtering relevant information from Twitter relying on emergency broadcasting services was proposed (Abel et al., 2012). As a new reported incident is detected, such information is parsed into a query for collecting tweets related to the incident. Each collected tweet is then processed by a semantic enrichment module based on NER techniques and using the DBPedia[41] and the OpenCalais for detecting persons, locations and organizations mentioned in a tweet. The Boilerpipe[42] system is used to explore the content of mentioned URLs.

A platform for disaster response in order to classify tweets during a crisis has also been proposed (Imran et al., 2014). The tweets could be collected using keywords or geographical boundaries of the target area. Human volunteers were used to perform two classification tasks over a sample dataset: the first one was to classify tweets into informative or not, and the second one was to classify the informative tweets into donation or not. Donation means messages related to donation actions performed during crisis. The CrowdFlower[43], a platform which assigns jobs to human volunteers who are paid for it, was used to perform those classifications tasks. The sample human-labeled dataset was used for training a classifier using the Random Forests algorithm in Weka[44] data mining system (Hall et al., 2009), and then to classify the whole dataset of tweets. The AUC (Area under an ROC curve) was chosen to measure the classification quality.

In the domain of forest fires, a prototype system to analyze social media content was proposed (Spinsanti and Ostermann, 2013). An automated classifier runs by a hand-crafted rule set based on the occurrence of some keywords, since Spinsanti and Ostermann (2013) had tried unsuccessfully using J48 and Naïve Bayes due to too much noise. The geocoding is performed by a simple string matching unigrams algorithm using a gazetteer containing names of provinces.

An ontology-based approach to automatically extract spatiotemporal and semantic information from news reports about hazards was proposed by Wang and Stewart (2015). They combined NLP and GIR techniques in order to show how using ontologies in GIR

[41] http://dbpedia.org/
[42] http://code.google.com/p/boilerpipe/
[43] http://www.crowdflower.com/
[44] http://www.cs.waikato.ac.nz/ml/weka/

applications is helpful for connecting multiple perspectives of hazards. The ontology was developed from authoritative data sources on hazards and by reusing concepts from existing ontologies such as OpenCyc[45] and GeoNames. The information extraction was performed using an extension of the GATE (General Architecture for Text Engineering) POS-Tagger, which combines three different gazetteers (spatial, temporal and semantic). The Yahoo's geocoding API was adopted as a geoparser. The authors discuss challenges related to complex phrases that involve vague location information and temporal references, and which remain open in the research community.

2.5.3 Ontology-driven *versus* Machine Learning algorithms for Information Extraction

The classical definition of ontology is provided by Gruber (1993): "An ontology is a formal explicit specification of a shared conceptualization". In other words, ontologies provide a shared vocabulary which can be used to model a domain. Such a shared vocabulary can be represented by objects and/or concepts that may contain properties and relations (Gruninger and Lee, 2002).

In computer science, ontologies have been used for the last decades for a set of tasks: improving communication between agents (human or software); enabling computational inference; and reusing data model or knowledge schema (Roussey et al., 2011). Ontologies can be classified according to the language expressivity and the scope.

According to the language expressivity (the knowledge representation language), there are the information ontologies, the linguistic ontologies, the software ontologies and the formal ontologies (Roussey et al., 2011). The information ontology is a clarified organization of ideas useful solely by humans. The linguistic ontologies can be dictionaries, folksonomies (Peters, 2009), lexical databases, etc (e.g. Resource Description Framework – RDF). The software ontologies focus on data storage and data manipulation for data consistency in software development activities (e.g. Unified Modeling Language – UML). The formal ontologies require a clear semantics and involve formal logic and formal semantics, with strict rules about how to define concepts and relationships (e.g. Web Ontology Language – OWL).

[45] http://sw.opencyc.org/

According to the scope, there are the domain ontologies, the core reference ontologies, the local ontologies, the general ontologies and the foundational ontologies (Roussey et al., 2011). The domain ontologies are only applicable to specific domains with specific viewpoints. The core reference ontologies are results of the integration of several domain ontologies, generally built to catch the central concepts and relations of a domain (e.g. CityGML[46]). The local ontologies are specializations of domain ontologies that represent particular models of domains according to a single viewpoint of a user (e.g. CContology[47]). The general ontologies contain general knowledge of a huge area (e.g. OpenCyc). Finally, the foundational ontologies are generic ontologies applicable to various domains that can be compared to the meta model of a conceptual schema (Fonseca et al., 2003) (e.g. Geography Markup Language – GML).

A way of extracting valuable information from texts is using ontology-based information extraction, as performed by Yang et al. (2012). Ontology-based IE processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information (Wimalasuriya and Dou, 2010). Two main categories split ontology-based IE: document-driven and ontology-driven. Document-driven (Corcho, 2006) is known as semantic annotation, where the discovered knowledge is structured in domain ontologies, while ontology-driven (Embley et al., 1998) extracts information from unstructured documents based on a domain ontology constructed under the help of domain experts or even by combining domain ontologies with core reference ontologies and folksonomies.

Most proposals from the state-of-the art information extraction from social media (e.g. Jin et al. (2013), Augustine et al. (2012), Yang et al. (2011)) rely on machine learning algorithms to build their classifiers such as Support Vector Machines (SVM) (Cortes and Vapnik, 1995), K-Nearest Neighbors (KNN) (Altman, 1992), Naïve Bayes (John and Langley, 1995), Multinomial Naïve Bayes (MNB) (Mccallum and Nigam, 1998), Bayesian Network (Friedman et al., 1997), J48 (Quinlan, 1993) and Random Forests (Breiman, 2001). Such proposals face problems concerning semantic knowledge representation. On the other hand, ontology-driven IE has emerged as a powerful approach for extracting spatiotemporal and thematic information in specific domains (Wang and Stewart, 2015).

[46] http://www.citygml.org/
[47] http://www.jarrar.info/CContology/

The output provided by both classification methods (ontology-driven or machine learning-based) are based on a set of classes according to each domain application. In the ontology-driven approach, such classes are defined in the domain ontology and returned at the final of the inference process. In the machine learning approach, such classes are provided in the training data.

One of the main advantages in using ontologies is the interoperability by humans and the ability of performing inferences, which makes the domain modeling easier and readable by humans. However, the performance of ontology-driven IE and classifiers based on most used machine learning techniques may vary according to the applied domain.

2.6 Summary

This chapter presented the basic terminology and the main concepts needed for understanding this research. Some LBSN for the smart cities domain are presented, such as the FixMyStreet, Crowd4City, ImproveMyCity, Colab, *ReclameAquiCidades*, BuitenBeter, *OndeFuiRoubado* and WikiCrimes. The Dublin Dashboard is presented to illustrate novel initiatives on cities where there are more than one LBSN addressing urban issues. The Geographical Information Retrieval (GIR) research field was presented together with the GeoSEn system for processing GIR. Concepts directly involved in this research such as Volunteered Geographical Information (VGI) and Ambient Geographical Information (AGI) were presented and compared each other. Finally, this chapter also presented an overview of approaches for information extraction on social media, the definition of ontologies in computer science and the parallel of ontology-driven and machine learning algorithms for information extraction.

The following chapter presents the related work on urban issue analysis from social media.

Chapter 3

Related Work

This chapter presents the related work on urban issues analysis from social media through four subsections. The first one describes the state-of-the-art on urban development projects based on social media data. The second one presents the efforts towards complaints identification from social media. The third one discusses existing proposals on location identification from social media. Finally, the last one provides a summary and a discussion comparing the strengths from both related work and this thesis.

3.1 Urban development projects based on social media data

The state-of-the-art in social media analytics on urban development projects has addressed many problems related to the urban environment and the citizens' behaviors. Such problems go beyond urban issues identification, as discussed below.

Xu et al. (2014) propose a method for real-time detection of urban emergency events from social networks, such as fires, storms and traffic jams. They use Weibo[48], a Chinese social network, by running a keyword-based search controlled manually with the aid of an API provided. The search results are manually labeled as positive or negative urban emergency event, based on accurate (e.g. "*I saw a car crash*") or inaccurate posts (e.g. "*I'm attending a crash conference*"), respectively. The method combines the

[48] http://weibo.com/

ICTCLAS[49], a Chinese part-of-speech tool, and the cosine similarity (Sidorov et al., 2014). The spatial information is detected and extracted using Baidu Map[50], which performs geocoding limited to China, Hong Kong, Macau and Taiwan. The temporal information is extracted directly from the post's timestamp.

Eshleman and Yang (2014) present a study which tries to extract the mood of Twitter by making a spatiotemporal association with an urban issues database available on the 311 civil-complaint service, from the city of San Francisco, USA. The authors are motivated by the fact that no studies on the relationships between Twitter data and social events available from government entities had been conducted since then. The DBScan algorithm is used to produce spatiotemporal clusters. The text content from both tweets and 311 service reports are not explored. Eshleman and Yang (2014) keep the focus solely on spatiotemporal aspects relying on tweet metadata.

Zheng et al. (2014b) focus on noise pollution, the third largest category of urban issues in the 311 civil-complaint service from New York City, USA. The distribution of noise pollution in New York City is analyzed by combining data from four different sources: 311 complaint data, user check-ins from social media, POIs and road network data. However, Zheng et al. (2014b) do not perform information extraction or detection of urban issues from social media.

Xia et al. (2015) propose a framework for real-time event detection from Instagram and Twitter post streams in the context of urban areas. They analyze several methods for integrating information from both social networks and discuss how the detection accuracy could be improved from such integration. Three main modules compose the framework: Event Signal Discovery, Event Signal Classification and Event Summarization. The discovery of events is based on both location, by means of small boxes from an urban area, and time series. The flow of tweets was continuously monitored and any abnormality was enough to produce an event candidate. Event candidates are then classified as either true or false events based on the extracted features: social, topical, emotional and spatial. True detected events are then summarized by analyzing images, topics, location and time. Several experiments are performed by using a dataset of geotagged tweets and photos captured during a 1.5-year period in the New York City.

[49] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html
[50] http://map.baidu.com/

Some journalists and trained users were recruited to produce a ground truth that could be used for evaluating the detection process.

Del Bimbo et al. (2014) present a Web GIS application which assembles clustering and visualization paradigms in order to provide exploration of urban social dynamics in terms of people lifestyle, business, demographics, transport, among others. Their main aim is to perform venue classification using semantics extracted from both places and online profiles of people who were frequent visitors of those places. They try to model Facebook user preferences in urban spaces by applying the k-means clustering algorithm on user check-ins and using Foursquare data to perform place categorization.

Lee et al. (2013) propose a model to characterize urban areas based on crowdsourced information available on Twitter. They collect a set of geotagged tweets and performed analysis in order to discover behavioral patterns in an urban area. The urban space is classified by crowd activities such as working, commuting, relaxing and eating/drinking. A monitoring system for harvesting twitter messages related to a geographical region of interest was developed. The geographical region of interest could be an urban area or even a whole country. Such system relied on simplified crowd behavioral logs, focusing on spatiotemporal metadata from tweets without deeply analyzing text messages. Three variables are considered within the model for each space-time span: the total number of shared messages; the number of different users; and the number of different mobile users (i.e., users changing their geographic locations over time).

Ngo and Revesz (2011) combine crash reports data and snow complaints data from the city of Lincoln, USA, to design a GIS system that enables real-time visualization of snow-related complaints. They describe a geocoding algorithm used to move data from an older dataset to a new one.

Zheng et al. (2014a) define the urban computing concept through a connection among urban sensing, data management, data analytics, and service providing. Urban computing is an interdisciplinary field where computer sciences meet city-related fields in the context of urban spaces. According to the authors, urban computing is an emerging area which can contribute to greener and smarter cities. The conceptual article discusses key challenges, datasets frequently used, applications, usually employed methodologies

and future directions. According to Zheng et al. (2014a), travel recommendations and location choosing for a business are issues addressed currently by the community. However, there is no discussion concerning current work in progress on urban issues identification. Zheng et al. (2014a) highlight the lack of proposals to identify and estimate the impact of changes in the city space, which can be performed through the integration of data management, machine learning and intervention-based analysis.

Crooks et al. (2015) highlight the potential of crowdsourcing for the urban domain. They clearly define form and function of the urban area and establish relationships between them. While the form is equivalent to the geometry of a city, the function refers to activities that are taking place within the known space of a city. Based on urban form and function definitions, authoritative geographical data and crowdsourcing generated data are compared. Several examples of how to extract urban form and function from crowdsourcing are provided, such as road network derivation, spatial area thematic classification, and spatiotemporal sentiment mapping.

Crooks et al. (2015) conclude that there is a notable shift from authoritative to crowd-harvested content which will require a new paradigm of collecting, analyzing and modeling urban morphology. Basically, there are three key areas related to the usage of crowdsourcing in urban modeling: collecting and curating crowdsourcing data, analysis and visualization, and scale issues. Collecting and curating crowdsourcing data involves heterogeneity and big data due to numerous data sources and volumes. Analysis and visualization is other related area since there are implicit and explicit content and context to be handled. Finally, scale issues involves the correct definition of the appropriate analysis scale in terms of both time and space dimensions.

3.2 Complaints Identification from Social Media

This section highlights the state-of-the-art on complaints identification from social media. Some approaches to complaints identification from social media may be useful for automated identification of urban issues, as any specific proposal could not be found in the literature. The complaints identification from social media has been studied in the last years as a step forward the identification of complaints on structured texts, such as reports or data from specific information systems. Many researches have focused on the domain of

customer complaints. Such domain is characterized by users complaining about products or services that have a great interest of companies, as those complaints may affect the business negatively.

Jin et al. (2013) propose an approach to identifying customer complaints from social media based on supervised text mining. Such an approach relies on two datasets: complaint and non-complaint post examples. The non-complaint examples include advertisements, advices and experiences, for example. Those datasets were then used in an enlargement algorithm that identifies complaints and non-complaints from a third dataset of unlabeled posts. A classifier was constructed by combining these datasets using Support Vector Machines (SVM) and k-Nearest Neighbors (KNN) classification algorithms, focusing solely on the complaints for the specific domain.

According to Jin et al. (2013), there are two main different methodologies for performing such a kind of classification task. One methodology is based on the division of the training documents into three sets: a small set of positive examples, a small set of negative examples, and a large set of unlabeled examples. These sets are then used to build and refine the classifier. The other methodology employs two-step heuristics. At the first step, negative examples are extracted from the unlabeled examples. At the second step, the classifier is then built using these negative examples together with a set of given positive examples. The authors commented that their approach can efficiently distinguish complaints from non-complaints in social media, especially when the number of labeled samples is very small. Their results also have shown that SVM has presented better results when compared to KNN.

Augustine et al. (2012) describe a system for outage detection of online services such as Netflix. Such system, named SPOONS (Swift Perceptions of Online Negative Situations), works on tweets, focusing on specific keywords to classify tweets as an outage complaint for the specific service. The authors noticed that a certain percentage of Netflix customers have found social media, such as Twitter, as another outlet for their complaints. In outage events, Augustine et al. (2012) found that there is usually a distribution of complaint tweets along the time until the disruption finishes.

The SPOONS system contains a gatherer component that queries Twitter for all English tweets containing the word “Netflix” or the hashtag “#netflix”. First of all, the

system observes a normal tweet day-to-day traffic and establishes a distribution pattern. Then, the complaint detection only starts when an "abnormal" behavior is detected from the tweet time-series distribution. The detection method relies on the occurrence of text fragments such as "is down" and on the Exponential Smoothing algorithm (NIST/SEMATECH, 2012) for spike detection. The system was evaluated through an official list provided by Netflix about the outages that have occurred during a time period.

He et al. (2014) focus on the online consumers' complaints. Using Weibo Chinese social network for evaluation, an influence measurement model of complaint theme is proposed. Such model is based on three factors from social media posts: the complaint text's quality, which relies on the length of the posts and keywords frequency; the user interaction degree, which analyzes quantitatively the different users who share a complaint; and the transmission timeliness, which is the forwarding frequency of a complaint during a certain period of time. The applied methodology focuses on keywords from complaints without considering spatial or temporal terms in the text. The identification of complaints from social media is based on previously-defined phrases that compound queries performed on the social media data.

Yang et al. (2012) focus on the food safety domain with the proposal of a domain ontology for hazard information extraction from Chinese food complaint documents. In order to provide a system able to automatically extract food complaints and produce early warnings, the authors rely on combining seed and related words to get the actual meaning of the complaint document in a semantic level. The system performs ontology-driven IE, which extracts information from unstructured documents based on a domain ontology. The seed words are the hazard information picked out from the food ontology. The related words of seed words are provided by word clustering using the Hownet[51], an online knowledge base that enables to perform the calculation of word similarity between Chinese words.

Three extraction steps are considered in the Yang et al. (2012) work: the extraction of background knowledge, the extraction of negative information and the extraction of hazard information. A method proposed by Stanford University is used for the ontology construction, while the Protégé ontology editor is used for modeling. The

[51] http://www.keenage.com/zhiwang/e_zhiwang_r.html

authors did not provide details or references about the method used for the ontology construction. The Jena Framework[52] is adopted for parsing the ontology in the application. Although the proposed method seems to present a good performance, the semantic reasoning ability of the proposed method is limited, since it is purely based on bag of words and no more than 10 related words are produced for each seed word.

Yang et al. (2011) propose a monitoring system by combining information retrieval techniques and a partially supervised learning algorithm to identify customer complaint posts on social media. Basically, the partially supervised learning tries to extract positive and negative examples from an unlabeled dataset aiming at augmenting the training data. The authors argue that their approach is an alternative to systems that label a large number of training data to be used in a supervised learning approach. Indeed, customer complaints on social media are highly distributed and non-complaint posts are diverse in topics, which makes labeling costly and time consuming. Although the partially supervised learning approach is more efficient, it may introduce noise that will cause many false positives and negatives. As the classification relies on the training data to perform good results, such noise can considerably decrease the system performance.

Ahltorp et al. (2014) trained a Conditional Random Fields model to detect medical complaints in a set of Japanese health reports. They propose a medical complaint extraction system which applies machine learning for recognizing complaints in texts and rule-based knowledge for detecting the modality of the recognized complaints.

Nakhasi et al. (2012) developed a preliminary study on Twitter aiming at identifying complaints in the health domain (complaints reporting medical errors). Tweets related to patient safety were identified and characterized through a procedure based on specialists reviewing a set of tweets manually. Those tweets were collected using specific key phrases, e.g. “sue the doctor” or “nurse screwed up”, on queries delivered to the Twitter API during a three-month period. Such study highlighted that patients are expressing their patient safety experiences on Twitter. However, no automated approach to the identification of complaints is presented.

Lee et al. (2015) combine domain ontologies and case-based reasoning for customer complaint handling. Focusing on the domain of restaurants, they developed and

[52] https://jena.apache.org/

applied an ontology for modeling knowledge and vocabularies. The proposed ontology was built by analyzing complaint documents and applied on cases retrieval and indexing. The main aim is to explore the nature of customer complaints and, subsequently, build a complaint-handling process. The proposed work focused on the semantic of complaints in such a specific domain. However, the ontology-based approach seems to be promising when used for complaint identification, since the particularities of specific domains are properly addressed.

Zirtiloğlu and Yolum (2008) propose a complaint management system which relies on an own developed ontology. This ontology focuses on urban issues specifically related to noise and traffic issue types. It was developed to provide a structured vocabulary and constraints for citizens who act as users of management systems to register their urban complaints. This work do not involve social media neither information extraction. It consists of an approach based on citizen complaints that were commonly made in person, by telephone or post at that time.

3.3 Location Identification from Social Media

There are many researches addressing the power of location-related social media. Magdy et al. (2014) present a system for scalable querying, analyzing and visualizing geotagged tweets relying exclusively on tweet metadata. Neither geoparsing nor geotagging techniques are applied by them. Hawelka et al. (2014) focus on human mobility through the analysis of geotagged tweets containing explicit geographic coordinates. They could show the power of geotagged tweets discovering human behaviors based on both space and time, although geotagged tweets represent just a tiny portion of all Twitter messages.

Yin et al. (2014) present an algorithm for toponym resolution and definition of the locational focus of tweets. Such algorithm relies on three information sources: body of the message, hashtags and the user location. The StanfordNER tool retrained using tweets (Liang et al., 2013) was used to perform toponym recognition aided with a simple hashtag segmentation method and replacing abbreviations with their respective full names through a dictionary. The main issue from such algorithm is that it relies on supervised learning,

which leads to the problem of provide a training dataset with enough manually annotated tweets.

While most authors focus on the location provided on tweet metadata, others investigate the identification of geographic locations in social media messages focusing exclusively on the text. Jung (2011) presents a method of analyzing sets of tweets aiming at identifying contextual clusters of tweets. By establishing a contextual relation between the messages, a set of tweets can be considered as a single document and make the process easier for the geoparsers. This task, however, can be very costly, depending on the volume of related tweets. In addition, there is also a possibility of errors in the geographic precision.

Watanabe et al. (2011) propose an automatic method of identifying geographic locations in non-geotagged tweets. Such method is based on the clustering of messages according to the type of event, considering short time intervals, small geographic areas and geotagged tweets. Geotagged tweets are then used to allocate geotags in tweets which do not have the geographic tag yet. The authors do not consider the possibility of the geotagged tweets having a different geographic reference than the location discussed in the messages. In addition, users do not necessarily talk about their current locations. Therefore, there is a possibility of errors in the geographic precision and this must be considered.

A general issue on proposals that aim at processing tweets to identify mentioned locations is the GLoD in which the geoparser is able to resolve. The higher GLoD enabled in most of current geoparsing techniques is City, which means that current geoparsers can only identify toponyms related to city names. In the context of smart cities, such geoparsers cannot be useful since toponyms inside urban areas, such as Districts, Streets and POIs, are not addressed. As gazetteers are the main data source of geoparsers, the GLoD provided by a geoparser reflects such a GLoD enabled in its gazetteers. Therefore, an approach to gazetteer enrichment is proposed in this thesis to enable geoparsers to resolve toponyms at a higher GLoD.

3.3.1 Gazetteer Enrichment for Toponym Resolution in Urban Areas

The next generation of gazetteers was prophesied by Keβler et al. (2009) few years after the OSM become popular. They linked the geographical information contribution and retrieval with the aim of using VGI in order to enrich gazetteers, and

discussed challenges to make it real. In their opinion, the next generation gazetteer needs to cope with harvesting and integration of information, assessing fitness for purpose, and enabling retrieval, querying and navigation. The key topic is the time-dynamic behavior of spatial features facing a world where things change quickly and that goes against the idea of static gazetteers.

The current literature comprises some proposals concerning gazetteers and leveraging VGI sources for making gazetteers up to date. Lamprianidis et al. (2014) proposed a method for extraction and integration of POIs from several crowdsourced media, e.g. DBPedia, OSM, Wikimapia and Foursquare. The main idea is to classify such POIs by a common taxonomy in order to gather POIs from different sources into a unique database.

Cardoso et al. (2016) propose a new generation of gazetteers using VGI and Semantic Web tools, such as ontologies and Linked Open Data. They presented the Semantic Web Interactive Gazetteer (SWI) and demonstrate that it can be used to add absent geographic coordinates to biodiversity records. SWI has a GeoSPARQL endpoint that enables complex semantic queries. However, even proposing gazetteer enrichment, their focus is on records based on municipality-level toponyms.

Gelernter et al. (2013) addressed gazetteer enrichment motivated by the fact that GeoNames is not rich in local toponyms. They proposed a fuzzy SVM-based algorithm that checks both OSM and Wikimapia for approximate spelling and approximate geocoding. Such a proposal compared the same toponym retrieved from both VGI datasets to reduce the noise and establish reliability on enrichment data to be integrated into the gazetteer. They suggested that OSM and Wikimapia would make suitable sources for gazetteer enrichment although the OSM coordinates have proved to be more geographically accurate.

Moura and Davis Jr. (2014) propose a gazetteer enriched with semantic relationships and connections with non-geo entities through combining GeoNames and DBPedia. Such work was motivated by the fact that there was not a single gazetteer that provides detailed coverage of places as intra-urban places. Their experiments show DBPedia has more urban details than GeoNames. However, neither of them has much information at such level of coverage.

Beard (2012) proposes an ontology-based gazetteer model for organizing VGI. The idea of the proposed ontology is to provide an alignment between VGI from different sources in order to enrich a gazetteer. Such enrichment can be done by two different ways: enriching previous stored features into gazetteer or creating new features.

3.4 Summary

By exhaustively reviewing the literature, it was possible to realize that no research addressing the automated identification of urban issues have been found in the literature. However, the relationship between the urban computing and social media data has been explored in other ways such as urban emergency events, people humor and lifestyle, noise pollution and mobility issues. The urban computing is an emerging research field, as stated by Zheng et al. (2014a), with many problems to be addressed, including urban issues which may help the improvement of cities infrastructure and the quality of services provided by local governments.

Table 3.1 presents a comparative among the related work and the proposal of this thesis for the automated identification of urban issues. Although it may appear a comparison among works with different purposes, the focus for such comparison is on the strategies applied for extracting information from the thematic, spatial and temporal facets. For such, related work in event detection and characterization for urban areas are compared along with proposals for extracting complaints from social networks, since they can provide relevant insights for working with the urban issues domain. Thus, such comparison is based on six relevant features:

- Data: the nature of the data explored by the work, such as crowdsourced (social media) or traditional (documents, tech reports);
- Language: the text languages supported by the work;
- Thematic facet: how the thematic facet is handled, if applicable;
- Geographical facet: how the geographical facet is handled, if applicable;
- Temporal facet: how the temporal facet is handled, if applicable;
- Performance: the statistical metrics for classifiers performance such as accuracy, precision and recall.

Table 3.1: Comparative table among related work and the proposal of this thesis

Work	Main Goal	Data	Language	Thematic Facet	Geographical Facet	Temporal Facet	Performance
Xu et al. (2014)	Detection of urban emergency events	Crowdsourced (Weibo)	Chinese	Keyword-based + lexical analysis + POS-Tagging	Geocoding limited to China, Hong Kong, Macau and Taiwan	Post's timestamp	Average precision = 0.95
Xia et al. (2015)	Event detection in urban areas	Crowdsourced (Instagram & Twitter)	English	Keyword-based + abnormal behavior monitoring	Only previously geocoded data used	Post's timestamp	F-Score = 0.84
Lee et al. (2013)	Characterize urban areas based on online activity	Crowdsourced (Twitter)	Japanese	Keyword-based without analyzing textual messages	Only previously geocoded data used	Post's timestamp	Unavailable
Jin et al. (2013)	Identifying customer complaints	Crowdsourced (HanTing club)	Chinese	SVM + KNN	Not addressed	Post's timestamp	F-Score = 0.65
Augustine et al. (2012)	Outage detection of online services	Crowdsourced (Twitter)	English	Keyword-based + abnormal behavior monitoring + J48	Not addressed	Post's timestamp	F-Score = 0.65
Yang et al. (2012)	Extract Chinese food complaints	Traditional (documents)	Chinese	Ontology-driven IE + domain ontology (words)	Not addressed	Not addressed	F-Score = 0.96
Yang et al. (2011)	Identifying customer complaints	Crowdsourced (Wal-Mart forum)	English	n-grams + clustering	Not addressed	Not addressed	F-Score = 0.70
Ahltorp et al. (2014)	Detecting medical complaints	Traditional (health reports)	Japanese	CRF model + rules-based knowledge	Not addressed	Not addressed	F-Score = 0.83
This Thesis	**Urban issues identification and classification**	**Crowdsourced (Twitter and LBSNs)**	**English**	**Ontology-driven IE + domain ontology (words and terms)**	**Geoparsing toponyms inside urban areas (in *non-geocoded* tweets)**	**Temporal Tagger + post's timestamp**	**F-Score = 0.69**

As it can be noticed, both the temporal and geographical facets are minimally addressed in the major of the proposals from related work. These works focus mainly on the thematic facet by adopting different techniques that include supervised algorithms from machine learning, keyword and rules-based reasoning and ontology-driven IE. The complaints identification in texts is a challenging task, which basically relies on natural language processing, machine learning techniques and in a well modeled knowledge of the domain. In addition, the nature of social media data poses an additional complexity as the texts tend to be shorter, sparse and present imprecise language, for example.

Focusing on the domain of urban issues, an additional challenging aspect is the need for managing the geographical facet. Urban issues are related to specific venues in an urban space that are usually mentioned in the text as the issues are reported. The detection of toponyms in texts such as social media posts is a bottom line for discovering useful spatially-related information such as complaints regarding issues on urban areas. However, in addition to dealing with social media data, the geoparsers need to be improved in order to be able to resolve toponyms inside urban areas.

Moreover, the work developed by Xia et al. (2015) cannot be applied to the urban issues identification as such issues are not reported like an upcoming event in the urban area. They are sparse on time, nature and geographical location. The observation of a considerable amount of people reporting the same urban issue in a relatively short period of time tends to be rare.

The urban issues analysis from social media data plays an important role for the improvement of the citizens' life quality, as the social media has proved to be a huge source for mining several aspects of population behavior. Crowdsourced information can provide a huge and cheaper overview of the main issues claimed, which may enable the government entities to provide effective solutions. Thus, on the one hand, this research aims to close the gap in the state-of-the-art on thematic identification by developing a novel approach with focus on social media data and addressing mainly the thematic and the geographical facets, and minimally the temporal facet. On the other hand, this research aims at filling the gap in the state-of-the-art on urban computing by providing a novel way of extracting urban issues that could be managed by citizens and local authorities. In the following chapter, the proposed solution for urban issues identification from social media is described in details.

Chapter 4

Identifying Urban Issues from Social Media

This chapter presents the proposed approach to identifying urban issues from social media.

Chapter 3 presented the main open problems concerning urban issues identification. Two problems are of particular interest in this work: the lack of approaches to model the urban issues domain; and the lack of approaches that address both thematic and geographical facets for the identification of urban issues from crowdsourced data. In order to overcome such questions, an approach to the identification of urban issues on social media is proposed considering mainly the thematic and the geographical facets, and minimally the temporal facet.

This chapter is structured as follows. Section 4.1 presents a formal definition of urban issues extracted from social media streams. Section 4.2 presents the details of the Urban Issues Domain Ontology (UIDO). Section 4.3 provides the details regarding the proposed approach concerning the thematic, geographical and temporal facets of urban issues. Finally, section 4.4 provides a summary of this chapter.

4.1 A Formal Definition of Urban Issues from Social Media

In order to build an ontology that models the domain of urban issues reported on social media, this section presents a formal definition of urban issues and discusses the main facets and concepts related to this domain. The section starts from raw social media messages (or just *posts*) and clearly specifies the facets of urban issues from user complaints shared on social media.

In social media, an urban issue (ui) rises from an urban issue report (r_{ui}) found in a post shared by a social network user. The entire social media post or just a fragment can be assigned to an urban issue report. Within urban issue reports, social media users may be complaining regarding an unsolved issue or thanking someone who acted on solving it. There may be many different reports regarding the same urban issue. The number of reports about a specific urban issue (n) varies according to different aspects, such as population, amount of online citizens, location, and time, among others. Thus, a set of urban issue reports (R_{ui}) regarding an urban issue is represented as $R_{ui} = \{ r_{ui_n} \mid n \geq 1 \}$.

The semantics of an urban issue report (r_{ui}) can be expressed by the quintuple presented in Equation 4.1.

$$r_{ui} = (user, description, issue_{type}, time, location) \quad (4.1)$$

Where:

- $user$ is the social media user who shared the report through a post;
- $description$ is the report description. It can be assigned to the entire social media post or just a fragment;
- $issue_{type}$ is the classification for the reported urban issue;
- $time$ is the timestamp for the report. The timestamp in which the report was posted or a time expression mentioned in the description can be assigned to the report time;
- $location$ is the place related to the reported issue. The geocoded location attached to the social media post or a place name mentioned in the description can be assigned to the report location.

Looking into an urban issue report (r_{ui}) structure, it can be noticed that there are three main facets in the urban issue semantics: thematic (*what*), geographical (*where*) and temporal (*when*). The thematic facet relates to the urban issue essence and its classification: $description$ and $issue_{type}$. The geographical facet relates to the urban space where the issue occurs: $location$. The temporal facet may relate to an approximate timestamp when the issue occurred: $time$ because urban issue reports are written in a variable time after the problem actually starts and the user may not be explicit on describing an exact time span. Finally, there is also the $user$ who wrote and shared the r_{ui}.

There may be many different issue types that classify urban issues. "Road Problems", "Rubbish" and "Leaks" are examples of issues types. It is important to provide a classification of urban issues as it is related to the hierarchical organization of cities into administrative authorities (or *city councils*) acting in specific urban areas. Each city administrative authority can be in charge of specific issues in the city and the number of such authorities may vary according to the city as well as the number of urban issue types.

The number of urban issue types is an open question: to the best of my knowledge there is neither a fixed number of issue types nor an official classification list for urban issues. The current instance of FixMyStreet in the Republic of Ireland, for example, contains six urban issue types, whilst the current instance running in the UK contains more than twenty. Other example is the San Francisco 311 service, which contains ten main urban issue types and many subtypes. Hence, each LBSN in the smart cities domain contains its own issue type set. Considering a hypothetical urban issue classification consisting of a number urban issue types, the issue type set (I) is formally defined as $I = \{ issue_{type_{\mathrm{m}}} \mid m \geq 1 \}$.

The impact that an urban issue provides on an urban area is related to the number of reports about such an issue. Therefore, the same urban issue (ui) can be reported (r_{ui}) by a set of bothered social media users located at the same or very close location in an urban area, in a temporal nearness and combined with the same $issue_{type}$: $issueType(r_{ui_a}) = issueType(r_{ui_b})$. The spatial nearness between different reports is identified based on a given radius between their geographical coordinate: $spatialNearness(r_{ui_a}, r_{ui_b}, radius)$. The temporal nearness between different reports is identified based on a given time span: $temporalNearness(r_{ui_a}, r_{ui_b}, time\_span)$.

Therefore, the issue type ($issue_{type}$), the geographical location ($location$) and the time span ($time$) are crucial to characterize the set (R_{ui}) to which each urban issue report (r_{ui}) belongs to. This makes the issue type as the main factor in the thematic facet of an urban issue report. Therefore, each urban issue report (r_{ui}) from an issue report set (R_{ui}) must satisfy the following constraint (Equation 4.2):

$$\forall\, (r_{ui} \in R_{ui})\ \exists\, (r_{ui'} \neq r_{ui}) \quad | \quad \left[\begin{array}{c} spatialNearness(r_{ui}, r_{ui'}, radius) \ \wedge\ temporalNearness(r_{ui}, r_{ui'}, time\_span) \\ \wedge\ \ issueType(r_{ui}) = issueType(r_{ui'}) \end{array} \right] \tag{4.2}$$

Given R_{ui}, Equation 4.1 and the stated constraint for R_{ui} members, an urban issue (ui) is then defined by the Equation 4.3.

$$ui \ \equiv\ (\ R_{ui}\ ) \ \equiv\ (\ issue_{type}, location, \{(time, user, description)\}) \tag{4.3}$$

Therefore, an urban issue from social media data is equivalent to a triple consisting of the issue type, the location and a subset composed by another triple composed by descriptions shared by social media users and related to specific time.

Having defined the concept of urban issues from social media, it is important to discuss how urban issue reports will be semantically identified and classified through a domain ontology. Hence, this section also defines the vocabulary for urban issues reported on social media and establishes links between vocabulary terms and urban issue types.

The vocabulary consists of a set of words (W) (or *lemmas*) in a particular language, such as English, which are commonly used by users reporting urban issues on social media. The number of words in such vocabulary is not constant, as the vocabulary may expand as new relevant words are discovered. Therefore, the set of words for the urban issue domain (W_{ui}) is formally defined as $W_{ui} = \{\ w_{ui_x} | \ x\ \geq 1\ \}$.

The vocabulary also contains a set of terms (T) where each term (t) is composed by one or more words from W. Thus, the set of terms for the urban issue domain (T_{ui}) is formally defined as $T_{ui} = \{\ t_{ui} \mid\ t_{ui}\ \equiv\ \mathrm{W} \subseteq W_{ui}, \mathrm{W}\ \neq\ \emptyset\ \}$.

Finally, each issue type in the urban issues domain contains a set of terms ($T_{issue_{type}}$) which is a subset of T_{ui}, as defined by Equation 4.4.

$$\forall(issue_{type} \in I)\ \exists(T_{issue_{type}} \subseteq T_{ui}) \quad (4.4)$$

As stated in Equation 4.2, the issue type, the location and the time are the main factors in an urban issue report. Similar to the urban issue types, a location can be classified in several levels, such as "Country", "City" or "Street". Moreover, a location level is part of the semantics regarding a location. It is needed to take into account a set of location levels ($L \neq \emptyset$), mainly the ones inside urban areas (e.g. district, street and POI), that classify each location. The number of location levels may vary for each country due to they are driven by administrative boundaries defined by local governments.

Given that specific urban issues may be related to particular location levels, a semantic relationship was also established between urban issue type terms and location levels. For example, the term "broken footpath" from the issue type "Road/Path Defects" has more sense if associated with locations at "street" and "POI" levels. An association between such a term with location levels "city" and "district" would be inaccurate and probably useless in the context of urban issue reports.

As urban issue types are identified and classified according to their set of terms defined by Equation 4.4, each issue type term must be associated with a subset of location levels (L') as defined by Equation 4.5.

$$\forall\left(t_{ui} \in T_{issue_{type}}\right)\ \exists(L' \subseteq L, L' \neq \emptyset) \quad (4.5)$$

The main concepts related to the taxonomy of urban issues and to the automated extraction of urban issues from social media streams were defined. The following sections describe the details of the proposed approach to urban issues identification from social media, including the ontological modeling of the concepts and relations defined in this section.

4.2 The Urban Issues Domain Ontology

Ontologies can be developed from reusing existing ontologies or just from a set of concepts still not modeled on other ones (Fernández-López and Gómez-Pérez, 2002; Gruninger and Lee, 2002). According to Simperl (2009), reusing ontologies can reduce the development costs because it avoids re-implementing components already available. The intended usage of the ontology will help defining the most suitable approach for its development.

There are many approaches for ontology design (Haghighi et al., 2013) such as the based on induction, deduction, inspiration, synthesis and collaboration. The design approaches that well suit the Urban Issues Domain Ontology (UIDO) are synthesis and deduction because the learning process is based on a corpus of urban issue reports. However, as ontologies are adaptable and can be refined, the collaboration and inspiration approaches can also be considered in the future through improvements by domain specialists and other people interested on enriching the ontology. Based on methodologies for building ontologies, Figure 4.1 presents the process flow for the development of the UIDO ontology.

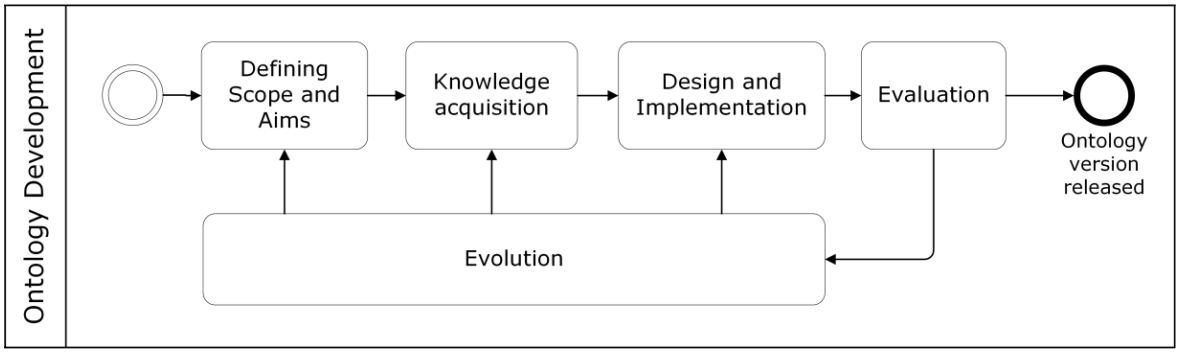


Figure 4.1: The UIDO ontology development process flow

The five steps of the ontology development are explained in the following subsections.

4.2.1 Defining Scope and Aims

The UIDO ontology aims at supporting semantic information extraction and classification in the domain of urban issues. For such, the UIDO ontology needs to model the specific vocabulary used on reporting urban issues. Moreover, it is important to define

a taxonomy of urban issues, named issue types, as discussed in Section 4.1. Such a taxonomy is useful for the LBSN in the smart cities domain because it enables LBSN users to search for specific urban issues besides the location constraints. Furthermore, semantic relationships between location levels and urban issue types are modeled aiming at supporting the automated identification of urban issues from social media.

Table 4.1 and Table 4.2 present sample reports in the domain of urban issues intended to be modeled by the UIDO ontology. From the examples shown in both tables, the UIDO ontology intends to cover issue type, vocabulary and location level fields (see Equation 4.3). The scope of the UIDO ontology encompasses all the concepts, constraints and relationships presented and discussed in the formal definition of urban issues.

Table 4.1: An example of an issue related to rubbish in an urban area

Report:	"Shameful mound of litter in Huguenot graveyard on Merrion Row. [URL]"
Source:	Twitter
Issue Type:	"Rubbish or Illegal Dumping"
Vocabulary:	"mound of litter"
Location level:	street: "Merrion Row" POI: "Huguenot graveyard"
Coordinates:	53.338506, -6.254964

Table 4.2: An example of an issue related to leaks in an urban area

Report:	"water leak Dublin Road, Skerries for 3 days opposite Texaco"
Source:	Twitter
Issue Type:	"Leaks and Drainage"
Vocabulary:	"water leak"
Location level:	street: "Dublin Road, Skerries" POI: "opposite Texaco"
Coordinates:	53.576952, -6.112755

In order to develop the ontology, this research relies on the FixMyStreet LBSN running in the UK and in the Republic of Ireland, due to the available crowdsourced datasets in the urban issues written in English. In addition, FixMyStreet was chosen because it is continuously used by the population on both urban areas, consisting of a relevant and up-to-date crowdsourced dataset in the domain of urban issues.

The current scope of the UIDO ontology consists of:

- Six urban issue types, that can be applied to many urban areas worldwide; however, some issue types being reported on social media may not be modeled by the ontology yet.
- A fixed number of vocabulary terms, as the research focuses on data from particular geographical area and culture.
- A single language addressed by the vocabulary: **English**; according to the available datasets.
- Three location levels inside urban area: "**district**", "**street**" and "**POI**". Although there may be other levels inside certain urban areas worldwide, these three are certainly found in most of them.

A secondary aim on developing the UIDO ontology is to establish a standard taxonomy of urban issues for LBSNs in the domain of smart cities in order to provide interoperability among them and, consequently, the reuse of information by other systems. Thus, the information regarding urban issues can be much more relevant to be used in city planning approaches. Although this research has based on FixMyStreet data, the UIDO ontology can be learnt and enriched with data from other datasets or with knowledge from domain specialists.

4.2.2 Knowledge Acquisition

The two main approaches for knowledge acquisition in ontology engineering are based on specialist knowledge and on corpora (Kang et al., 2014). The UIDO ontology learns trough an engineering process based on corpora. This process is similar to the processes adopted by Wanner et al. (2015), Amini et al. (2015) and Onorati et al. (2014).

Such corpora consist of two FixMyStreet report datasets acquired along the entire 2015 (around 18,000 English-written urban issues reports). The details regarding such datasets are given in Section 5.1 (Chapter 5).

Instead of searching for specialists, developing questionnaires and then analyzing them, the knowledge acquisition process focuses on the behavior of people who use the LBSN to report their complaints regarding urban issues. That is, the complainant citizens act as a kind of specialist in a given domain. Therefore, this process adopts an approach based on "the voice of the crowds" within a crowdsourcing environment (Surowiecki, 2005).

In order to model the urban issues vocabulary in the UIDO ontology, the six urban issue types from the FixMyStreet datasets (Graffiti, Leaks and Drainage, Litter and Illegal Dumping, Road or Path defects, Street Lighting and Tree and Grass maintenance) are considered subdomains of urban issues. Such an assumption is needed to correctly identify relevant terms for each specific subdomain aiming at identifying urban issues and performing the classification through one of the issue types available. Thus, the ontology learning process relies on six datasets, each one related to an urban issue type, useful to discover subdomain relevant terms.

To date, fully-automated ontology learning and engineering is an open research field. There are some proposals which help the process of knowledge acquisition from corpora (Cimiano and Völker, 2005; El-Beltagy and Rafea, 2009; Tonelli et al., 2011; Kang et al., 2014); nevertheless, the majority of this process remains manual and human-dependent due to the challenge to automatically define and connect concepts correctly in a particular domain.

As there are corpora available for learning and a common knowledge about urban issue features, an ad-hoc approach was developed to extract the main concepts and terms for the urban issues domain. This approach is based on the partially-automated method for key concept extraction proposed by Kang et al. (2014), which involves noun phrase extraction, stopword removal, synonym finding and candidate enrichment. Therefore, the approach comprises a statistically-based heuristic to extract the most relevant terms of the urban issues domain. Figure 4.2 shows the process flow of the developed statistically-based heuristic in detail.

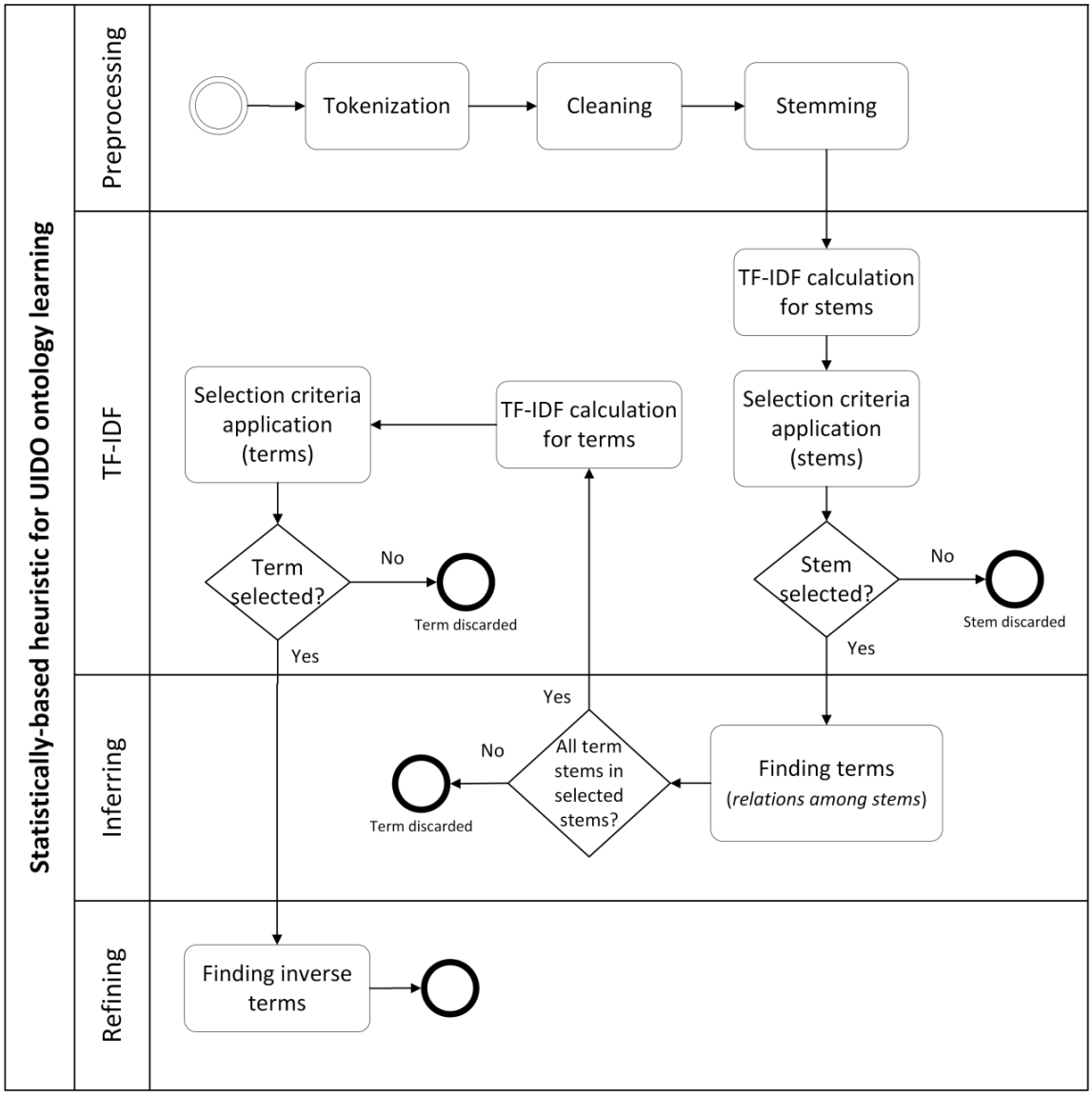


Figure 4.2: The process flow of the statistically-based heuristic developed

The automated process for the UIDO ontology learning comprises nine steps in four distinct stages (preprocessing, tf-idf, inferring and refining) organized as shown in Figure 4.2. Such process starts with a corpus to be analyzed and finishes with a set of relevant terms and their stems to be included in the UIDO ontology as instances. Each one of the six issue type datasets available for learning was processed.

The preprocessing stage comprises three steps: tokenization, cleaning and stemming. The reports are tokenized and a bag of words is created with distinct words from the corpus. The bag of words is cleaned in the cleaning step, which removes ordinal number suffixes, numbers as tokens and months of the year. Then, the stemming step finds

the root of words in order to reduce derived words to the word stem (e.g. light replaces lighting and lights). For this, the Porter's stemmer algorithm (Porter, 1997) was used. Notice that the preprocessing stage for the UIDO ontology knowledge learning does not explore stop word removal as preliminary experiments demonstrated that some stop words may be used in relevant terms for urban issues identification. The remaining useless stopwords are removed automatically in further stages. Listing 4.1 presents a summarized Java algorithm that illustrates how the preprocessing stage proceeds. The preprocessing stage finishes forwarding the stemmed corpus and the bag of stems to the tf-idf stage.

Listing 4.1: The summarized Java algorithm for the preprocessing stage

```
List<Report> reports = readFromFixMyStreetDataSet( issueType );
List<Word> bagOfStems = pre_processing( reports ) {
        List<Word> bagOfWords = tokenization( reports );
        bagOfWords = cleaning( bagOfWords );
        bagOfWords = stemming( bagOfWords );
        return bagOfWords;
};
```

In the tf-idf stage, the calculation of the term frequency-inverse document frequency (*tf-idf*) statistic (Spärck Jones, 1972) is performed for each one of the stems in the bag of stems produced in the previous stage. The *tf-idf* is widely used in Information Retrieval and Text Mining because its value reflects how important a word is to a document in a corpus. It is assumed that each urban issue report is as a document and each set of reports from an urban issue type is as a corpus. Listing 4.2 presents a summarized Java algorithm that illustrates how the tf-idf stage proceeds.

For each stem, three statistic parameters are calculated: the term frequency (*tf*), the inverse document frequency (*idf*) and the *tf-idf*. The double normalization *K* scheme with $K = 0$ (Wu et al., 2008) was adopted for the *tf* calculation. Such schema normalizes the variation on the number of stems among the reports. In order to calculate the normalized *tf* considering the entire corpus, the *tf* calculation is the average of *tf* in each report, as shown in Equation 4.6. N is the number of reports in the corpus.

Listing 4.2: The summarized Java algorithm for the tf-idf stage

```
function List<Term> tfidf_stage( List<Word> bagOfStems ) {
    Double tf_threshould = 0.04;
    Double tfidf_threshould = 0.12;
    bagOfStems = calculateTfIdf( bagOfStems );
    List<Word> reducedBagOfStems = wordSelectionCriteria(
                    bagOfStems, tf_threshould, tfidf_threshould);
    List<Term> bagOfTerms = inferring_stage ( reports,
                                            reducedBagOfStems );
    bagOfTerms = calculateTfIdfTerms( bagOfTerms );
    List<Term> reducedBagOfTerms = termSelectionCriteria(
                    bagOfTerms, tf_threshould, tfidf_threshould);
    return reducedBagOfTerms;
};
```

For the *idf* calculation, the normal scheme (Metzler, 2011), widely used by the community, was adopted, as presented in Equation 4.7. Finally, the *tf-idf* is simply the math multiplication $tf \times idf$.

$$tf_{stem} = \frac{\sum_1^N \frac{(\ number\ of\ stem\ occurrences\ in\ the\ report\ )}{(\ number\ of\ stems\ in\ the\ report\ )}}{N} \quad (4.6)$$

$$idf_{stem} = \log_e \frac{N}{(\ number\ of\ reports\ in\ the\ corpus\ with\ the\ stem\ in\ it\ )} \quad (4.7)$$

The stems with their statistic parameters values are submitted to a selection criteria developed based on both *tf* and *tf-idf* values. The selection criteria were defined in order to keep only relevant candidate stems from the analyzed corpus. Such criteria is based on the *tf X tf-idf* scatter plot (Trim, 2013), which explores the relationship between both statistic values to classify each word as stop word, frequent word or rare word.

Figure 4.3 shows an adapted *tf* X *tf-idf* scatter plot where the Y-axis is the *tf* and the X-axis is the *tf-idf*. This adaptation was needed because each corpus has different

quantities of reports and distinct stems. As the original *tf X tf-idf* scatter plot (Trim, 2013) does not clarify the boundaries of the word classification areas, approximate boundaries were defined through a preliminary analysis over a sample word set from the available corpora.

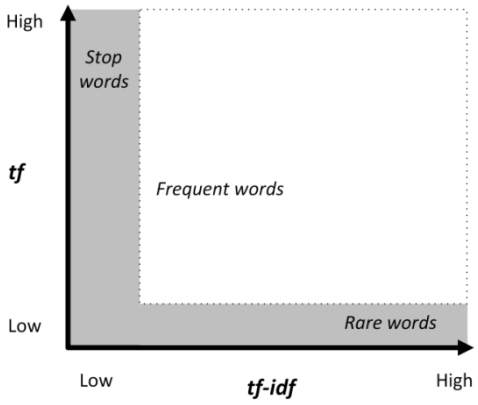


Figure 4.3: The *tf X tf-idf* scatter plot and word classification areas (adapted from Trim, 2013)

The UIDO ontology learning process is interested on discovering frequent words that can be relevant for the urban issues domain and their respective urban issue types. Although rare words are generally important for query analysis, they are not in this particular case. A preliminary analysis in a sample set of words classified as rare words has shown that most of them relate semantically with specific place names which were mentioned in some reported urban issues.

In order to select only frequent words that may be relevant for the urban issues domain, thresholds were defined for both *tf* and *tf-idf* values. This means that only stems with the pair (*tf-idf* , *tf*) greater than or equal the threshold are considered relevant and then selected in this step. As the number of reports in each corpus varies considerably (see Table 4.3), relative thresholds were defined for *tf* and *tf-idf*, that may vary according to the corpus size. To define such thresholds, the maximum *tf* and *tf-idf* values are calculated from the corpus. Thus, the *tf* threshold is 4% of the maximum *tf*, while the *tf-idf* threshold is 12% of the maximum *tf-idf*. Such threshold percentages were defined based on preliminary tests with different threshold percentages, manually visualizing the set of selected stems and identifying whether known relevant stems were being filtered out.

Figure 4.4 shows six scatter plots that visually help to understand the relationship between the *tf* (Y-axis) and *tf-idf* (X-axis) values from a stem, and how such relationship

can be helpful on identifying relevant and irrelevant words from a corpus. The points in the clear areas represent selected stems.

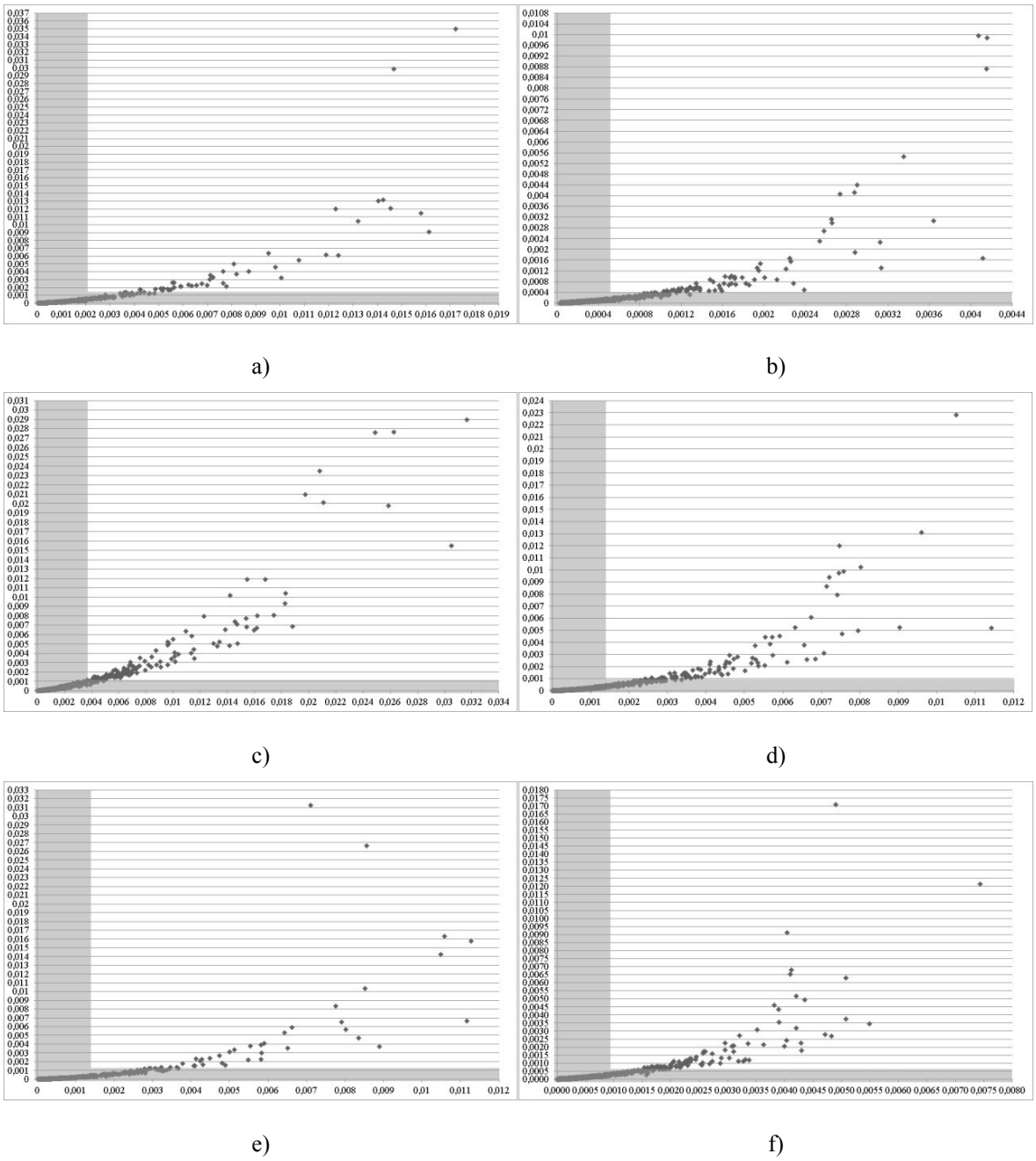


Figure 4.4: The *tf X tf-idf* scatter plots with the distribution of stems in each datasets: a) Graffiti, b) Leaks and Drainage, c) Litter and Illegal Dumping, d) Road or Path defects, e) Street Lighting and f) Tree and Grass maintenance reports dataset. The points in the clear areas represent selected stems.

Table 4.3 summarizes the quantity of distinct stems found in each corpus, as well as the quantity of selected stems and the thresholds applied for the stem selection.

Table 4.3: Data from the corpora after the selection criteria application in stems

Urban Issue Type	Reports	Stems	Selected stems	Reduction	*tf* threshold	*tf-idf* threshold
Graffiti	1,189	2,296	45	98.04%	0.0014003	0.0020667
Leaks and Drainage	259	1,324	70	94.71%	0.0003982	0.0004993
Litter and Illegal Dumping	8,878	8,649	125	98.55%	0.0011598	0.0037985
Road or Path defects	5,173	9,328	93	99.00%	0.0009138	0.0013707
Street Lighting	1,477	3,303	41	98.76%	0.0012518	0.0013532
Tree and Grass maintenance	1,154	3,502	88	97.49%	0.0006854	0.0008931

From the results shown in Table 4.3, it can be noticed that the developed selection criteria reduced the bag of stems to 2.24% of its original size in average. Continuing the process flow presented in Figure 4.2, the reduced bag of stems (for each corpus) is forwarded to the inferring stage.

The inferring stage comprises finding terms in the corpus related to the reduced bag of stems. In order to discover the frequent terms that may be useful in the urban issues subdomain, this research relied on the Text2Onto system (Cimiano and Völker, 2005). Text2Onto analyzes text documents and searches for relevant concepts and relations that would be relevant for an ontological model of the corpus. Such system relies on GATE (Cunningham et al., 2013), an open source solution for text processing, and WordNet (Miller, 1995), a lexical database for the English language. Thus, Text2Onto can run seven algorithms which aid to identify concepts, instances, relations, subclasses and subtopics from a corpus.

An algorithm was implemented using the Text2Onto library for Java[53], which sends the corpus to be processed by Text2Onto, parses the output provided in a Probabilistic Ontology Model (POM) and then analyzes the terms found. In the Text2Onto system, the POM represents learned knowledge at a meta-level in the form of instantiated modeling primitives (Cimiano and Völker, 2005).

[53] Text2Onto library for Java available from https://code.google.com/p/text2onto/downloads/list (last access in July, 2016)

Figure 4.5 illustrates the processing in the inferring stage through the algorithm developed. The Text2Onto output (POM) is translated in a bag of terms. As the inferring stage is interested on word-word relationships, all distinct terms (instances with two or more words) found in the POM were parsed and then proceeded by stemming and comparing each term with the reduced bag of stems. Only terms from the POM in which all their stems are contained in the reduced bag of stems comprise the bag of terms. This resulting bag of terms is then forwarded to the next step.

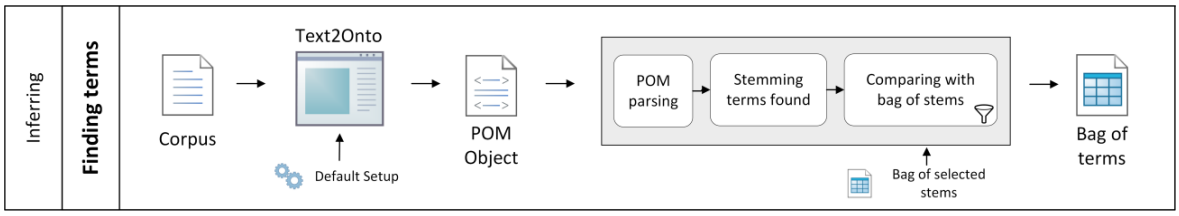


Figure 4.5: Details of the finding terms step in the inferring stage

Similarly as performed with the bag of stems extracted from the corpus, the selection of relevant terms was performed from the bag of terms returned by the finding terms step. The *tf*, *idf* and *tf-idf* for each term are calculated and then forwarded to the selection criteria application for terms (see Figure 4.2). The *tf* is calculated according to the Equation 4.6, however, as a term has more than one stem, the number of stems in the report is normalized considering the number of stems in the term (Equation 4.8). N is the number of reports in the corpus. The *idf* is calculated based on Equation 4.7 but changing the context for terms instead of stems (Equation 4.9). Finally, the *tf-idf* is the multiplication $tf \times idf$.

$$tf_{term} = \frac{\sum_{1}^{N} \frac{(\,number\ of\ term\ occurrences\ in\ the\ report\,)}{(\,number\ of\ stems\ in\ the\ report\,)}}{(\,number\ of\ stems\ in\ the\ term\,)} \qquad (4.8)$$

$$idf_{term} = \log_e \frac{N}{(\,number\ of\ reports\ in\ the\ corpus\ with\ the\ term\ in\ it\,)} \qquad (4.9)$$

The terms with their statistic parameters values were submitted to the selection criteria step developed based on both *tf* and *tf-idf* values, similarly to the process adopted

for stems (see Figure 4.2). Table 4.4 summarizes the quantity of distinct terms found in each corpus, as well as the quantity of selected terms and stems and the thresholds applied to the term selection.

Table 4.4: Data from the corpora after the selection criteria application in terms

Urban Issue Type	Selected stems	Terms	Selected terms	Reduction	*tf* threshold	*tf-idf* threshold
Graffiti	45	35	20	42.86%	0.0014004	0.0020667
Leaks and Drainage	70	29	25	13.79%	0.0003983	0.0004993
Litter and Illegal Dumping	125	343	64	81.34%	0.0011599	0.0037985
Road or Path defects	93	184	56	69.57%	0.0009138	0.0013707
Street Lighting	41	52	37	28.85%	0.0012519	0.0013533
Tree and Grass maintenance	88	80	36	55.00%	0.0006854	0.0008931

From the results shown in Table 4.4, we can notice that the developed selection criteria reduced the bag of terms to 48.57% in average. Continuing the process flow presented in Figure 4.2, the bag of terms containing only selected terms (for each corpus) is forwarded to the refining stage.

The refining stage comprises the step for finding inverse terms, the last step in the proposed heuristics for extracting relevant terms for the urban issues domain. The finding inverse terms step consists of searching the corpus for occurrences of inverse terms from the bag of selected terms. Inverse terms are terms that contain the same set of stems but in different order. Semantically, inverse terms may be synonym terms. For example, searching the corpus for inverse terms of the term “street light broken” from the bag of selected terms may return as result: “broken street light”. Inverse terms that are not already in the bag of selected terms are then included. This step is included in the heuristics as there might be inverse terms that are relevant to the domain but could not be included in the bag of select terms because they did not match the selection criteria. As inverse terms are usually synonyms, they are less frequently used and consequently they have lower *tf* and *tf-idf*.

After completing the learning process for each one of the six urban issue types, there are six bags of selected terms that model such urban issue types (see $T_{issue_{type}}$ and

Equation 4.4). These selected terms comprise our current model for the urban issue domain (elements from W_{ui} and T_{ui}). Such model is preliminary because it is based solely on a specific scenario. However, other scenarios may be further addressed in order to enrich this model. Figure 4.6 shows a word cloud with the stems from the term sets extracted and selected from the FixMyStreet corpora. This word cloud was generated using the WordItOut[54] cloud generator.

Figure 4.6: Word cloud of stems from the urban issues domain

Although such presented automated process was applied in a specific domain for the UIDO ontology learning, such process can easily be applied to other domains or even to other corpora in the domain of urban issues.

4.2.3 Design and Implementation

The design phase for the UIDO ontology started by analyzing the CContology (Jarrar, 2009), a domain ontology developed to support online customer complaint management. Such ontology models seven main concepts: Complaint Problems, Complaint Resolutions, Complaint, Complainant, Complaint-Recipient, Address, and Contract. From these ones, only four concepts would be relevant on modeling the UIDO ontology:

[54] http://worditout.com/

Complaint, Complainant, Complaint Problems and Address. However, these four concepts could not be directly reused because they focus on specific customer complaints that are quite different from the complaints in the urban issues domain. Therefore, only the CContology design ideas were used for defining the concepts and relationships for the UIDO ontology.

The UIDO ontology was developed in the Web Ontology Language (OWL) (McGuinness and van Harmelen, 2004) using the Protégé 5.0 system[55]. The core of the UIDO ontology is the concept of an urban issue report. The core class has two main goals: to act as a top concept during the reasoning process on discovering urban issues from a text; and also to be a metamodel for the Linked Data entities from the social media messages related to urban issues. Figure 4.7 shows the main concepts and relationships of the UIDO ontology. These concepts and relationships form the basic skeleton, but there are others in the sub-levels of the hierarchy.

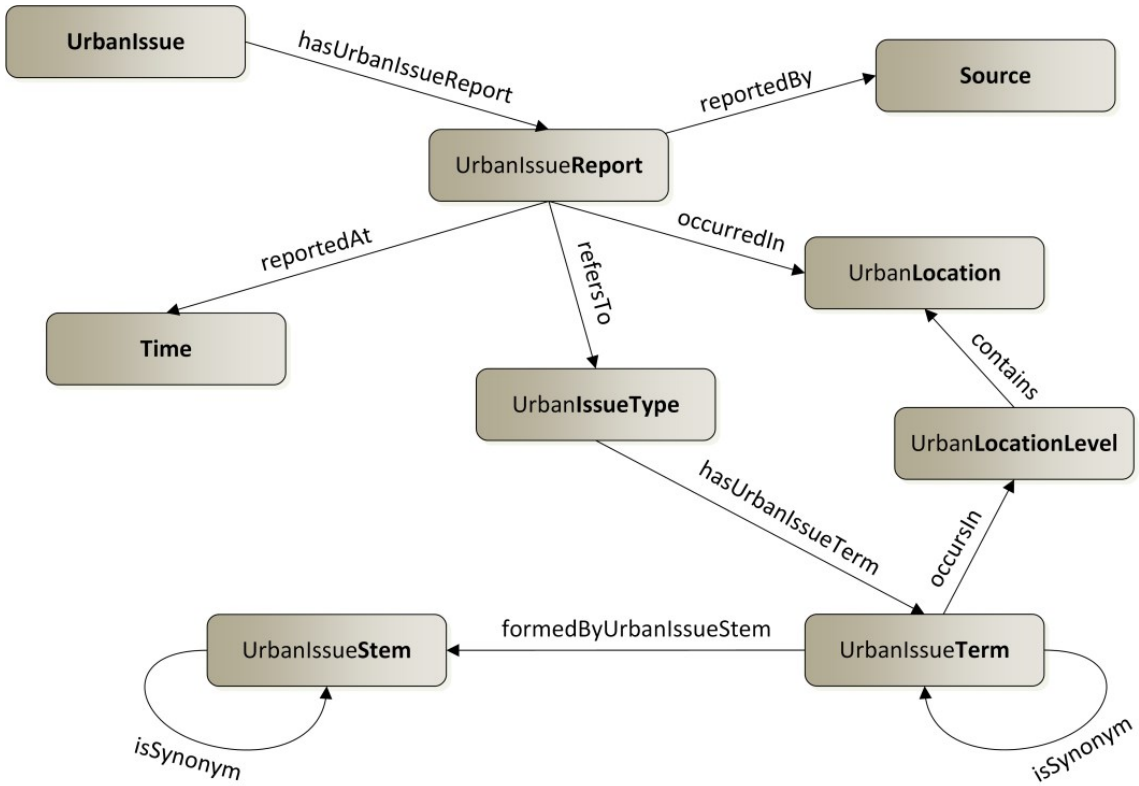


Figure 4.7: Main concepts and relationships from the UIDO ontology

[55] http://protege.stanford.edu/

The concepts urban issue type and urban location level (shown in Figure 4.7) are explained in more detail because they are the most important concepts beyond the core. The core class `UrbanIssueReport` models the reports regarding urban issues. The structure of the concept follows the Equation 4.1. The "user" attribute in Equation 4.1 is encapsulated into the `Source` class as well as the "description" attribute. The class `UrbanIssue` models the Equation 4.3, grouping urban issue reports from the same issue type, location and time by the relationship *`hasUrbanIssueReport`*, following the constraint stated in Equation 4.2.

The class `UrbanIssueType` models the classification of issues related to the urban environment ($issue_{type} \in I$). There are initially five subclasses for urban issue types: education, security, infrastructure, health and transportation. Figure 4.8 shows the urban issue type hierarchy. The six issue types learned from the FixMyStreet corpora fits in the infrastructure concept. Thus, they are included as `Infrastructure` subclasses. Other corpora and specialist knowledge would be necessary in order to expand the ontology within other classes and subclasses that may be further included.

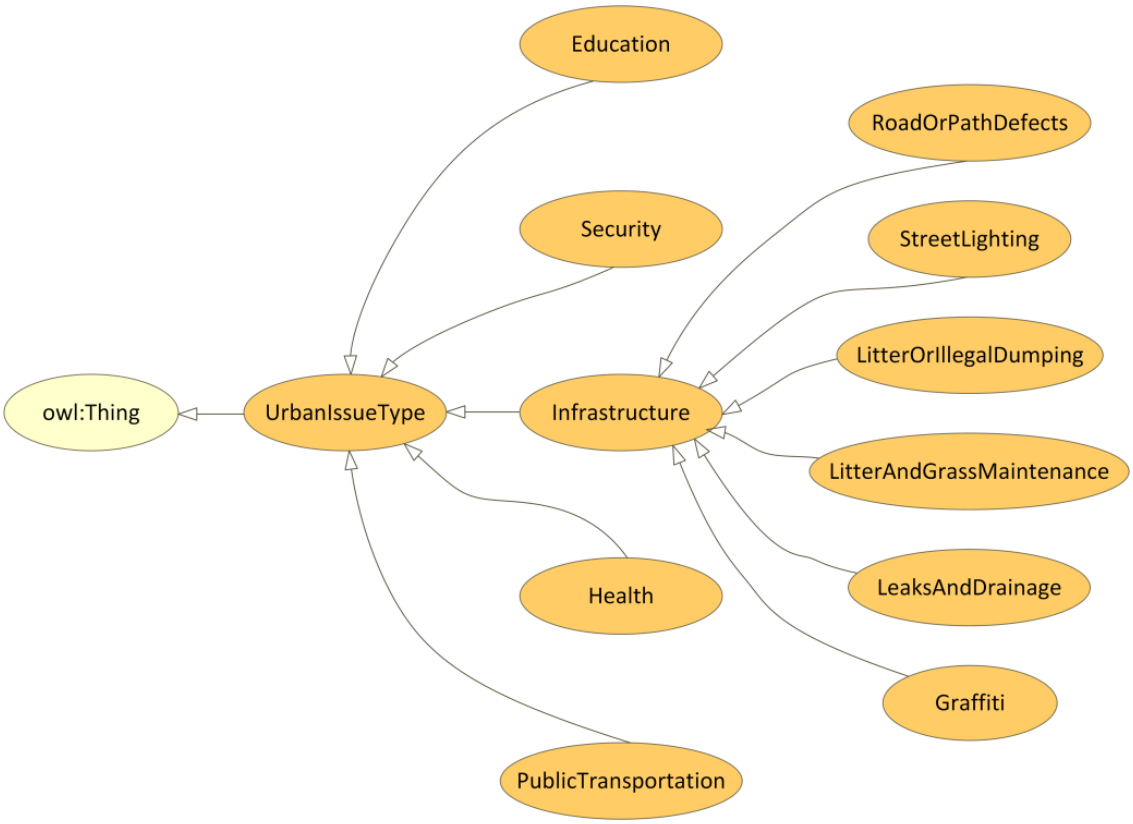


Figure 4.8: The urban issue type hierarchy

Figure 4.9 shows an example of two instances ("broken bollard" and "pavement damag") from the classes `UrbanIssueTerm` and `UrbanIssueStem` in a relationship with the issue type "Road or Path Defects" from the `UrbanIssueType` class.

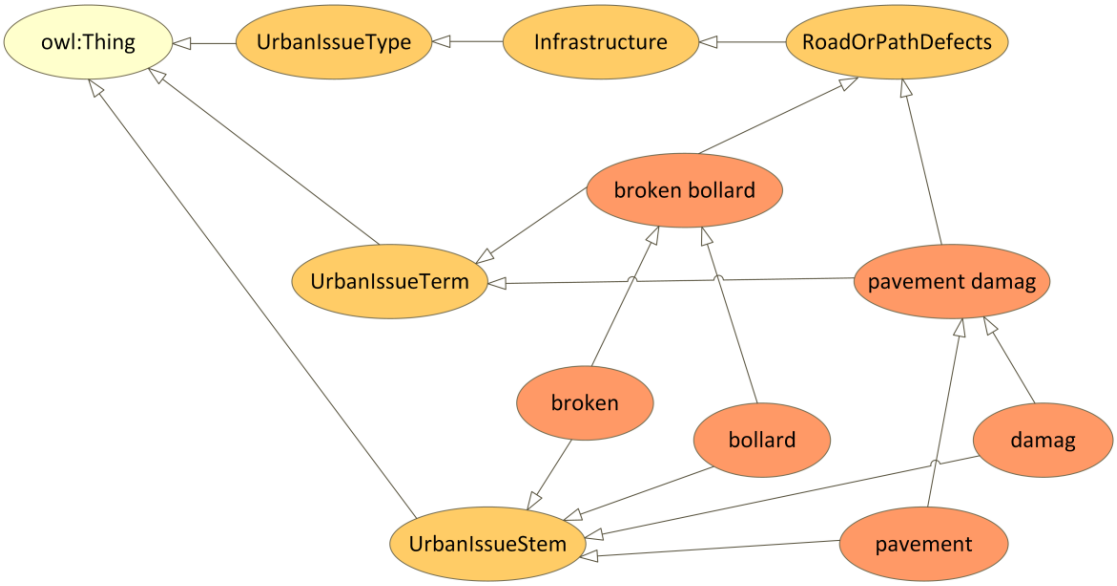


Figure 4.9: An example of urban issue stems and terms related to an urban issue type

The class `UrbanLocationLevel` models the level of detail for geographical locations. Figure 4.10 shows the urban location level hierarchy. Such levels are used in constraints for urban issues locations according to the urban issue term (Equation 4.5). These constraints are modeled through the relationship *occursIn*. For example, a detected urban issue report about a pothole is relevant if it occurred at street or POI levels, however, at lower levels such as city or county, such a report tends to be irrelevant due to the geographic vagueness. There are initially three instances: Street/Road, District and POI. However, other levels may be further included according to new studies. The semantics of such instances are explained in details in the following section.

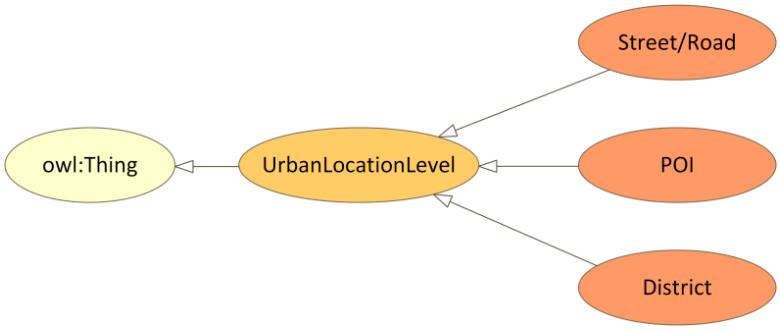


Figure 4.10: The urban location level hierarchy

The class `UrbanLocation` models the geographical information of urban issue reports. Such class reuses concepts from the LinkedGeoData ontology (LGDO) (Stadler et al., 2012). The LGDO ontology was chosen for the spatial locations related to urban issues for the following reasons:

1) It is derived from concepts defined by the OpenStreetMap (Haklay and Weber, 2008);
2) OpenStreetMap is currently a huge up-to-date and open spatial database that contains spatial features, mostly inside urban areas, such as street networks, which are commonly found in urban issue reports;
3) A preliminary work (presented in the following section) that focused on the geographical facet of urban issues enables to store spatial features from OpenStreetMap and provides the location levels for the UIDO ontology.

The class `Time` models the temporal information of an urban issue. The temporal information can be assigned to the time where the social media message was posted or it can be extracted from the social media message using temporal parsers. Such class reuses concepts from the OWL-Time ontology (OGC and W3C, 2016). The OWL-Time ontology models temporal concepts for describing the temporal properties of resources in the world or described in Web pages such as urban issues reported on social media. Thus, the time concept in the UIDO ontology is equivalent to the `TemporalEntity` in the OWL-Time.

Finally, the classes `UrbanIssueTerm` and `UrbanIssueStem` models the key terms learned during the knowledge acquisition stage. These concepts are in charge of connecting the top concepts modeled with the terms learned from the analyzed corpus. A set of relevant stems are instances of the urban issue stem concept instead of words (elements from W_{ui}) as a stem is the root of a word and thus it better represents group of words with similar semantics. The combined stems produce a set of relevant terms (T_{ui}) through the relationship *`formedByUrbanIssueStem`*. These terms are instances of the urban issue term concept. Then, subsets of these relevant terms compose the vocabulary of urban issue types (Equation 4.4) through the relationship *`hasUrbanIssueTerm`*. Both term and stems can be synonyms for each other. For those cases, the relationship *`isSynonym`* applies.

4.2.4 Evolution

The evolution stage can be achieved many times during the ontology development and implementation. With reference to Figure 4.1, the evolution may result in redefining scope and aims, in other knowledge acquisition stages or in adjustments in the implementation and design. In general, evolution in ontologies means refinements and knowledge expansion in such a specific domain.

For instance, evolution in the UIDO ontology may comprise the inclusion of other languages, the inclusion of other issue types in the infrastructure, or even the addition of other levels into the issue type's hierarchy. An evolution stage may involve learning corpora from other data sources or from an updated corpus from the FixMyStreet LBSN with reports from 2016. In addition to other corpora, stakeholders worldwide can be included in order to provide quality improvements. Moreover, the ontology may evolve from people feedback regarding classification of urban issues on social media posts.

4.3 The Proposed Approach

This section presents the proposed approach to the identification of urban complaints from social media. A systematic approach aiming at automatically producing AGI (Ambient Geographic Information) based on information published on crowdsourced data is proposed. The geographic information from social media posts combined with thematic from specific domains such as urban issues can be classified as AGI, a deviation from VGI as discussed in Section 2.3 (Chapter 2). Thus, social media users assume the role of volunteers in the production of AGI, which could automatically be made available to users of LBSN applications.

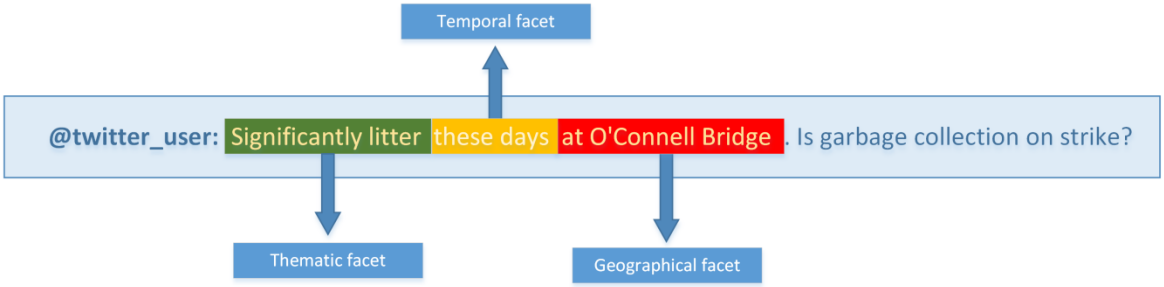
In an envisioned scenario, the produced AGI should become available for LBSN users, who will be the main consumers and also validators of that information, being capable of pointing to its inconsistencies as well as stressing its relevance, and consequently enriching the crowdsourcing environment. As the thematic focus is on urban issues, such produced AGI would be in the context of smart cities. The proposed approach also keeps the focus on the English language, as it can be applied in many urban areas worldwide since English is spoken in many countries worldwide.

Figure 4.11: The main idea of this proposal: turning social media messages into valuable AGI in a LBSN

An overview of the general idea of the proposed work is presented in Figure 4.11. The Crowd4City (Falcão, 2013) is used as an LBSN example, which can be automatically enriched through the proposed approach with AGI derived from social media, such as Facebook and Twitter. The scope of this research focuses solely on Twitter data. The Facebook screen shown in Figure 4.11 is only for illustration of the developed ideas. However, Facebook and other social media networks can be addressed in further work.

In the context illustrated in Figure 4.11, each tweet can be turned into AGI which can then be used by LSBN users. This research also limits the focus on the identification of urban issues because one of motivations is to enrich LBSNs in the smart cities domain with useful data automatically extracted from social media data. However, the proposed approach is generic as it can be used in other application domains.

In order to provide automated identification of urban issues that will be translated into AGI to be used in LBSN environments, the proposed approach relies on methods for detection and classification in the three facets of social media data: thematic, geographical and temporal. These facets have relationships in the urban issues domain, as discussed in Sections 4.1 and 4.2, which need to be taken into account. For example, an urban issue concerning potholes (*thematic*) is implicitly related to the urban area geography (*geographical*) such as the streets network or POIs, as well as such an urban issue is valid for a specific time span (*temporal*) because it is expected the pothole will be fixed in a further moment. An example of an urban issue reported on Twitter is illustrated in Figure 4.12.



Metadata

parameter	type	description
created_at	STRING	Time the tweet was posted (with timezone)
coordinates	FLOAT[]	Point location where the tweet was created
retweet_count	INTEGER	Number of times the tweet was retweeted
lang	STRING	BCP 47 language identifier detected from the tweet
hashtags	OBJECT[]	List of Hashtags mentioned in the tweet
user_mentions	OBJECT[]	List of Twitter users mentioned in the tweet
urls	OBJECT[]	List of URLs mentioned in the tweet
user_id	INTEGER	The identifier of the user who posted the tweet
user_location	STRING	The user-defined location
...	...	...

Figure 4.12: Example of an urban issue reported on Twitter and some useful metadata

The top part of Figure 4.12 highlights the three facets of an urban issue that can be found in the tweet message body. The green text relates to the thematic facet, while the red one relates to the geographical information and the yellow one to the temporal information. The main focus of the proposed approach is on the thematic and geographical facets. The inclusion of the temporal facet considering the challenging temporal aspects would considerably increase the complexity of such a proposal. Nonetheless, the temporal facet is minimally explored by relying on the `created_at` parameter from metadata and using a third-party temporal tagger for inferring an approximate time period for the urban issues identified.

Figure 4.12 also illustrates some relevant fields from tweet metadata[56]. These parameters could be useful when performing the automated identification of an urban issue from Twitter. The `lang` parameter is crucial as the parser is language sensitive. The parameter `coordinates` can be helpful for the analysis of the geographical facet because the Twitter user may be close to the urban issue location. The `retweet_count`

[56] The complete list of tweet metadata parameters is available in: https://dev.twitter.com/overview/api/tweets

parameter can be used on the calculation of the urban issue relevance as the tweet is retweeted more people are agreeing with the issue. The `hashtags`, `user_mentions` and `urls` parameters could be used in a preprocessing step in order to reduce the noise on text data to be analyzed. Finally, the user-related parameters could be useful for analyzing the user in terms of home location and searches on its timeline feed. A detailed example of a tweet in JSON format provided by the Twitter API is shown in details in Appendix B.

Aiming at providing automated identification of urban issues from Twitter, the proposed approach can be described as a generic process for AGI identification comprising the following tasks: harvesting, preprocessing, thematic analysis, spatial analysis, temporal analysis, summarization and AGI production. Such process is generic because it can be applied to other thematic domains. Figure 4.13 illustrates such a process and the main tasks, which are explained in detail in the following subsections.

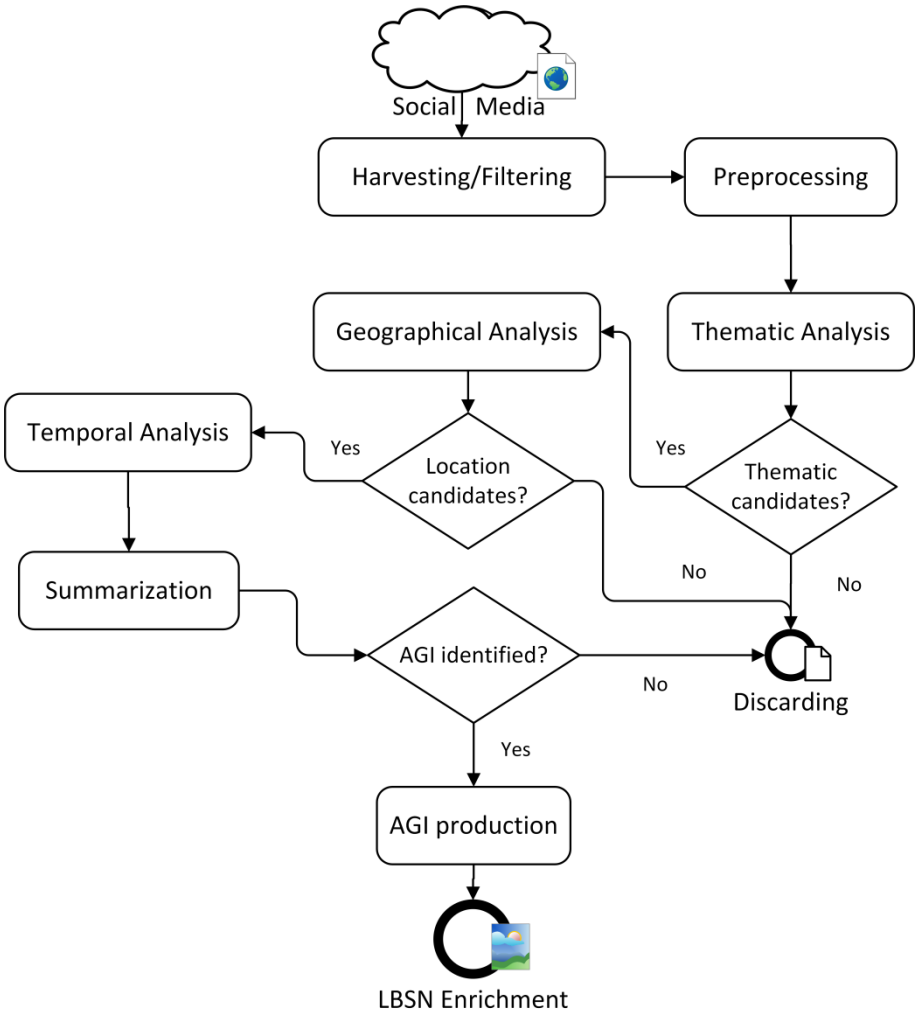


Figure 4.13: The generic process flow for automated identification of AGI from social media

In Figure 4.13, there is a conditional component regarding the identification of AGI from previous tasks. An AGI is identified if the summarization task resolved one thematic issue and one geographical location at least.

The process flow for automated identification of AGI illustrated in Figure 4.13 is suitable for the domain of urban issues because it comprises the three facets of the urban issues. However, the thematic analysis needs to focus on the specific domain in order to provide accurate AGI identification. Thus, using this generic process demand a work to be done on adapting the thematic analysis for the specific application domain regardless the adopted domain.

4.3.1 Harvesting/Filtering

Harvesting is the first task of the AGI identification. This task aims at monitoring the social media feeds in order to acquire data to be processed. As this research focuses on Twitter, it relies on Twitter Streaming API[57], which provides real-time monitoring on Twitter based on the language, a set of keywords and/or geographical boundaries through bounding boxes.

Considering that urban issues relates to the urban area of a city (and usually reported from a location within the same city), it is important to define the city that will be monitored. For this reason, tweets are retrieved through the Twitter Streaming API based on the geographical search. The amount of georeferenced tweets is still low (~1%) when compared to non-georeferenced tweets (Liu et al., 2014); however, it is more feasible to start looking at specific urban areas instead of just using a bag of keywords and retrieve spatially-sparse tweets.

Once the tweets that satisfy the geographical constraints were harvested, a filtering stage may be applied in order to filter out in advance the tweets which tend to be irrelevant (does not contain AGI) and reduce the dataset to be processed due to the fact that there may be thousands of new tweets harvested in a short time span. There are many ways of performing such filtering. One example is described in Appendix A, which details the harvesting and manual labeling process applied on tweets for the evaluation of this research.

[57] http://dev.twitter.com/overview/api

4.3.2 Preprocessing

Once a tweet is harvested, this is submitted to a preprocessing task, in which a preliminary analysis of the text is performed to reduce noise and prepare the text to be correctly analyzed in the following tasks.

Although tweets are known as microtexts due to the limitation of 140 characters in their body, some noise may be found. The tweet body can combine user mentions, URLs, hashtags, and the user message, which need to be correctly processed. The hashtags, for example, are usually seen as a combination of words without space that may be relevant on mining the text (Yin et al., 2014; Feyisetan et al., 2014; Lingad et al., 2013).

Furthermore, there may be particular noise in the tweet body as Twitter users are used to apply: informal language (e.g. misspellings, contractions and slangs); character repetitions; abbreviations due to the character-length limitation; special-characters combination to represent emotional symbols (*emoticons*); and imprecise description regarding the theme or location. Such issues are challenging and need to be addressed while dealing with social media data. Due the difficulties involved in addressing these issues, most recent researches have addressed them partially, according to its goals (Augustine et al., 2012; Abel et al., 2012; Imran et al., 2014; Yin et al., 2014; Feyisetan et al., 2014).

Unfortunately Twitter does not still perform NLP and NER to enrich tweet metadata with entities found in the text. Currently, NLP applied in the Twitter is restricted to the tweet language detection, which is attached to the tweet metadata. Therefore, in order to perform the preprocessing, the following preprocessing techniques are explored:

- **URL Removal:** removing URLs by replacing them with a special token like "_URL_", similarly as performed by Imran et al. (2014);
- **User mentions removal:** removing user names e.g. "@username" and the common retweet fragment "RT @username". This technique addresses privacy and accuracy issues, since such information is not relevant for the identification task;
- **Hashtag segmentation:** consists of making hashtags readable by semantic and geographic parsers. Similarly to Yin et al. (2014), this work opts for a simple hashtag segmentation method based on a word list to break hashtags. Such word list

is leveraged from both thematic vocabulary and geographical gazetteers. Then, it is possible to perform segmentation for words related to the application domain or a place name;

- **Character repetitions solving:** consists of reducing words with more than two sequentially-repeated characters. A simple algorithm to reduce such repetition into a single character is applied by following the approach carried out by Kouloumpis et al. (2011);
- **Stemming:** consists of finding the root of a word. Such technique decreases the number of unique words being evaluated. The original Porter's stemmer[58] (Porter, 1997) was adopted because it works with English language.

It is important to highlight that the preprocessing task does not exploit stop word removal as preliminary studies indicated that some stop words may be relevant for urban issues identification. More details regarding such aspect are described in the section related to the thematic analysis task. Hence, stop words are not prior removed and are dealt in each one of the facet analysis tasks.

Finally, a simply algorithm was developed to convert common abbreviations and informal terms or expressions into formal ones. Such algorithm relies on a list of formal terms/expressions and their respective informal variations. The main idea of using the conversion algorithm is to improve the performance of the stemming and hashtag segmentation algorithms, since they work better on formal structures of a language.

4.3.3 Thematic Analysis

The thematic analysis task focuses on identifying relevant terms for a specific application domain in order to produce the AGI. For such, this task is based on the ontology-driven Information Extraction. Ontology-driven IE seems to be feasible in this research because it enables performing inferences and relies on ontologies, which makes the domain modeling easier and readable by humans, unlike most used machine learning-based classifiers. Moreover, the work performed by Wang and Stewart (2015) demonstrated that the ontology-driven inferring is promising for extracting information

[58] http://tartarus.org/martin/PorterStemmer

from news reports. Thus, this work investigates the feasibility of using ontology-driven inferring for extracting information from social media data.

Instead of applying machine learning classifier algorithms, such as SVM and Naïve Bayes, this research relies on an approach based on knowledge structured in domain ontologies. While most used machine learning classifiers are limited in terms of semantic modeling and the generated models are difficult to be interpreted by humans, ontologies can model the domain semantics besides being interoperable by humans. The semantics are not limited to the domain essence due to the fact that there may exist relationships with geographical and temporal facets. Moreover, training with short texts such as social media is a challenge for machine learning-based methods due to the inadequate amount of data for the learning process.

The details of the thematic analysis described in this section are related to the urban issues domain, according to the scope of this research. Although other domains certainly require some adapting work for this task, some of the ideas discussed along this chapter may be applied on them.

A preliminary design for the thematic analysis task included sentiment analysis combined with ontology-driven inferring. However, a study carried out in a sample tweet dataset (see Appendix A) showed that the relationship between negative sentiment polarity and urban issue reports is not strong enough. Such finding suggested that the thematic analysis task should not rely on sentiment analysis for the identification of urban issues from social media.

Another line of research consisted of performing event detection from social media aiming at identifying urban issues. For such, the possible solutions are based on timeline analysis, which consists of continuously monitoring the stream in order to detect abnormal behaviors that may suggest natural hazards or upcoming events (Abel et al., 2012; Spinsanti and Ostermann, 2013; Imran et al., 2014). It could be noticed that such an approach based on event detection is not suitable to perform urban issues identification because the urban issues domain is not predictable as such kind of events. There is no specific time for an urban issue to be published and it is rare to have many people talking about the same urban issue. On the other hand, a domain ontology models the urban issues vocabulary and relevant terms relationship through ontological rules in order to be attached

in a thematic parser for the reasoning task on identifying and classifying urban issues reported on social media. Figure 4.14 illustrates how the thematic analysis proceeds.

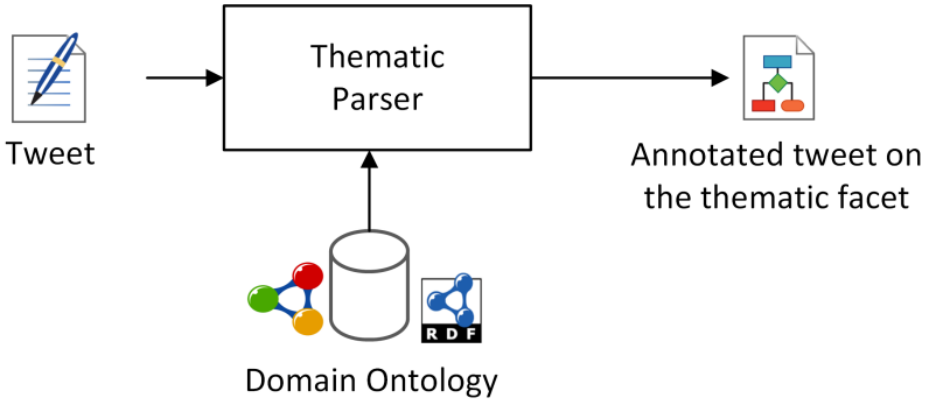


Figure 4.14: The thematic analysis sub process flow

The thematic parser is a software that processes the tweets in order to annotate them regarding the application domain. For such, a domain ontology is needed to provide the reasoning rules to be used in the reasoning process performed by such parser. To the best of my knowledge, a domain ontology in the urban issues domain comprising urban issue types was not proposed yet. Therefore, the Urban Issues Domain Ontology (UIDO) was proposed. The details of the design and development of the UIDO ontology is described in Section 4.2.

The thematic parser was developed to attach the UIDO ontology and to perform the reasoning regarding the identification of the thematic facet in processed data. Such parser is a module of the Social2AGI, a system prototype developed aiming at implementing the automated approach to the AGI production from social media presented in this chapter. The details of the Social2AGI system are presented in Appendix C.

The thematic parser is implemented in Java by using the Jena[59] open source Java framework to connect the domain ontology. The thematic parser functioning consists of checking the tokenized stems from a social media text in order to identify thematic candidate terms according to the inferring rules modeled in the domain ontology. Listing 4.3 presents a summarized Java algorithm that illustrates how the thematic parser proceeds.

[59] https://jena.apache.org/

Listing 4.3: The summarized Java algorithm for the thematic parser

```
List<SocialMediaText> msgs = getMessagesToParse();
DomainOntology onto = loadDomainOntology(UIDO.class);
for ( SocialMediaText msg : msgs ) {
    List<Word> stems = preprocessing(msg);
    List<ThematicCandidate> tcandidates = thematicParser(
                                                    onto, stems)
    {
        for ( Word stem : stems ) {
            if ( onto.modelsStem( stem ) ) {
                tcandidates.addAll(
                        onto.identifyCandidates( stem, stems ));
            }
        }
    };
}
```

The method `thematicParser(…)` receives the domain ontology (UIDO) and a list of stems from the message to be parsed. Such method connects to the UIDO ontology and scans the entire list of stems by searching for stems which are modeled by the domain ontology.

The boolean method `modelsStem(…)` from the domain ontology verifies whether a given stem is an instance from the class `UrbanIssueStem`. If this is the case, the method `identifyCandidates(…)`, also from the domain ontology, tries to identify urban issue type candidates starting from such instance of `UrbanIssueStem`. For such, the method first identifies the instances from the class `UrbanIssueTerm` which are `formedByUrbanIssueStem` (see Figure 4.7 for reference of UIDO classes and relationships). Then, `UrbanIssueType` subclasses are retrieved from those instances through the relationship `hasUrbanIssueTerm`. Table 4.5 shows an example of a tweet being processed by the thematic parser following the algorithm shown in Listing 4.3. Such example aids the understanding of such processing.

Table 4.5: Example of a tweet being processed by the thematic parser

Stage	Tweet
Got original message	`Huge potable water leaking near 142 O'Connell Street Upper Dublin`
Preprocessing	`huge potabl water leak near 142 o connel street upper dublin`
modelsStem('huge') = false	`huge potabl water leak near 142 o connel street upper dublin`
modelsStem('potabl') = false	`huge potabl water leak near 142 o connel street upper dublin`
modelsStem('water') = true	`huge potabl water leak near 142 o connel street upper dublin`
identifyCandidates(...)	`huge potabl water leak near 142 o connel street upper dublin` → **LeanksAndDrainage** → Infrastructure → UrbanIssueType (*relations given by the UIDO ontology*)
modelsStem('leak') = true	`huge potabl water leak near 142 o connel street upper dublin`
identifyCandidates(...)	`huge potabl water leak near 142 o connel street upper dublin` (*already included as thematic candidate*)
modelsStem('near') = false	`huge potabl water leak near 142 o connel street upper dublin`
...	...

In order to identify the `UrbanIssueTerm` instances involving a given stem, the method `identifyCandidates(…)` relies on the ordered list of stems from the message to be parsed. Thus, the method can scan the stems that appear before and after the given stem in the message aiming at performing matching among terms. Such term matching relies on the Jaro-Winkler distance (Winkler, 1990) to measure the similarity of the strings from the terms from the original message and terms modeled in the UIDO ontology as instances of the `UrbanIssueTerm` class. As the similarity score varies from 0 (*no similarity*) to 1 (*exact match*), the method matches terms with similarity above 0.8. Such threshold was defined after experimentations with some terms from the domain ontology and tweets.

After identifying thematic candidates, it is important to perform disambiguation of the terms by analyzing the word sense. Such task is complex, since word-sense disambiguation is an open problem of NLP. Thus, further work should investigate ways of enriching semantically the domain ontology in order to enable to perform term disambiguation.

Listing 4.4 shows an example of the JSON output from the thematic parser after the inferring process based on the domain ontology has finished. The attribute `tweet_id` is the real tweet identifier and is useful for retrieving all the tweet data in the subsequent tasks. The attribute `word_count` is the number of words found in the tweet, whilst the attribute `word_count_irrelevant` is the number of stop words found in the tweet.

Finally, the attribute `thematic` is an array with all the urban issues candidates found in the tweet. Such information is used in the summarization task, which is further explained.

Listing 4.4: Example of JSON output produced in the thematic analysis task

```
{
  "tweet_id": 651873880924143617,
  "word_count": 11,
  "word_count_irrelevant": 1,
  "thematic": [
    {
      "urban_issue_type": "leaks/drainage",
      "urban_issue_term": "water leak",
      "keywords": [
        {
          "word": "water",
          "position": "3"
        },
        {
          "word": "leak",
          "position": "4"
        }
      ]
    }
  ]
}
```

All the urban issues candidates found are sent to the subsequent task, the geographical analysis. A tweet that does not present any urban issue candidate has the `thematic` attribute equals to `NULL`. Those tweets are discarded and the other stages are aborted for such tweet. No AGI is produced on such case, as there is no thematic facet identified.

4.3.4 Geographical Analysis

The geographical analysis consists of detecting and classifying the references to place names (*toponyms*) eventually mentioned mainly in the body of a tweet. In the context of urban issue, such references represent the geographical facet of the urban issue report, being necessary for AGI production and LBSN enrichment. Therefore, the geographical analysis relies on the three location contexts that can be assigned to a tweet: the geocoded, the user home and the mentioned location.

The mentioned location context is related to the toponyms mentioned in the body of the tweets. On identifying urban issues reported in tweets, it is expected to find

mentions to the location related to the reported issue. In the manually labeled tweet datasets (see Appendix A), mentioned locations could be found in 55.2 % of the urban issues reported in Greater Dublin and 69.4 % of the urban issues reported in Greater London. Hence, the geographical analysis needs to rely on geoparser systems that enable to detect and resolve place names into geographical coordinates that could be then used for geographic referencing in the AGI produced.

There are a variety of geoparsers in the literature. Campelo and Baptista (2009), for example, proposed the GeoSEn system. The GeoSEn is described in detail in Section 2.2 (Chapter 2). An open issue with current open geoparsers is that their gazetteers (normally used as their main source of geographic data) do not cover geographical features at such high GLoD required in smart cities related approaches, such as POIs, Streets and Districts. As the geographical analysis needs to address locations inside an urban area, it is not possible to rely on a geoparser which has "city" as the higher GLoD. Faced with this issue, the GeoSEn system was extended in order to enable it to detect toponyms that refer to urban areas. The GeoSEn was chosen due to three main facts: 1) the geoparser is unsupervised, which means it does not need training; 2) the source code available and 3) the available direct support from the authors. The GeoSEn extension is explained in the following.

The GeoSEn system extension for English texts and urban area place names

The GeoSEn has two main limitations: it has the city level as the highest GLoD, similarly other gazetteers; and it is only enabled to handle texts written in Portuguese. Thus, an extension of the GeoSEn system was developed comprising the gazetteer enrichment, the ability of recognizing the English language, and the heuristics adjustments in order to enable the GeoSEn parser to perform toponym resolution for more geographically specific locations.

As discussed in Section 2.2 (Chapter 2), the GeoSEn gazetteer is composed of toponyms related to the Brazilian's political subdivision and its most geographically precise level for stored toponym is the city-scale level (see Figure 2.11 in Chapter 2). This may imply an inaccurate toponym resolution while geoparsing messages (e.g. tweets, etc.) related to a big city that may contain more precise information like districts or known buildings. This can cause too many toponym resolutions into the same location (typically

the centroid of that big city). The citizens interested on specific regions within their cities would need to manually filter all the information related to such cities, for instance.

Both the GeoSEn gazetteer and the GeoTree were extended in order to store and enable toponyms related to more specific locations within a city. The gazetteer enrichment focused on three types of toponyms: POI's, such as well-known buildings, shops or touristic targets; Street names and Districts/Boroughs. Such types of toponyms apply to both urban areas object of study in this research. Therefore, these types were organized and included into the previously established GeoTree. In addition, the lower levels have also been changed because it was originally developed for Brazilian toponyms. An example of an instance from the extended GeoTree is shown in Figure 4.15.

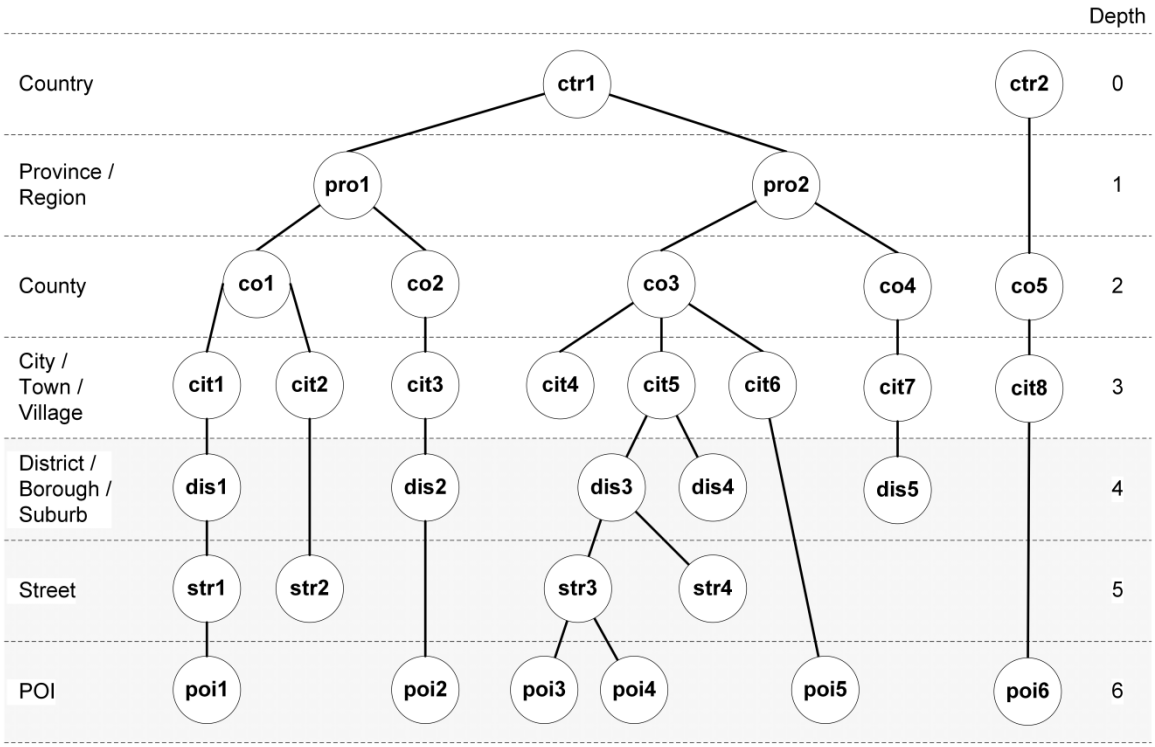


Figure 4.15: An example of an extended GeoTree instance

In the extended GeoTree shown in Figure 4.15, the most relevant levels for the urban areas are levels 4, 5 and 6. The main focus of the gazetteer enrichment is on those levels. In order to obtain the geographical data, the gazetteer enrichment relies on Volunteered Geographic Information (VGI), since it is continuously updated by volunteers widespread worldwide. Figure 4.16 presents the architecture for the gazetteer enrichment module.

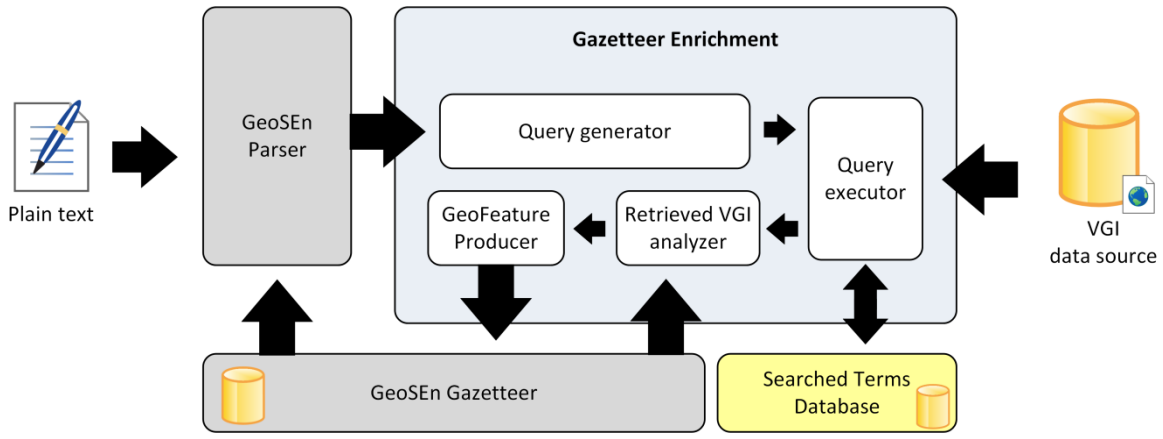


Figure 4.16: The gazetteer enrichment architecture

In the gazetteer enrichment method implementation, place name external searches are delivered on-demand as soon as a text is processed by the GeoSEn parser. The gazetteer enrichment method can be subdivided into four stages: query generation, query execution, retrieved VGI analysis and GeoFeature production. In the following, it is explained each stage considering tweets as the text data to be parsed.

- **Query generation:** performs the preprocessing over the text, which includes the removal of hashtags, URLs, "RT"s (a token commonly used in retweets[60]), usernames preceded by "@", and the generation of queries to be delivered to the query execution stage.
- **Query execution:** establishes the communication between the gazetteer enrichment method implementation and a VGI source using an API.
- **Retrieved VGI analysis:** performs the classification of the retrieved features and also seeks for duplicates into the GeoSEn gazetteer.
- **GeoFeature Production:** produces the record of new geographical features that will enrich the gazetteer and also produces adjustments into the previously stored features.

In order to store a high number of different toponyms into the GeoSEn gazetteer, several different search terms can be produced from a single tweet during the query generation stage. The basic idea of the algorithm is to combine between one and four

[60] A retweet is the name given to a tweet produced by sharing of other twitter previously published. A tweet can be retweeted many times by the owner or other Twitter users.

neighbor words to form a search term using white spaces between them. This strategy enables to get toponyms from a VGI source formed by just one and up to four words properly. Stop words, special characters and punctuation are discarded. Figure 4.17 illustrates some iterations of such algorithm in a sample text.

Tweet: **huge pothole near 490 leinster road rathmines**

Iteration	...	Status
i	huge pothole **near** 490 leinster road rathmines	*ignored*
i+1	huge pothole near **490** leinster road rathmines	*ignored*
i+2	huge pothole near 490 **leinster** road rathmines	*searched term*
i+3	huge pothole near 490 **leinster road** rathmines	*searched term*
i+4	huge pothole near 490 **leinster road rathmines**	*searched term*

Figure 4.17: Some iterations of the search-term-generator algorithm in a sample text

In the five iterations shown in Figure 4.17, it can be noticed that only three produced search terms delivered to VGI source. It is important to notice that, although the query "*leinster road*" should retrieve the same results as "*leinster road rathmines*" due to the first one being more generic, this is not guaranteed by the VGI source. Duplicate results eventually retrieved by a more specific search are discarded in the retrieved VGI analysis stage. The two iterations that have not produced and delivered a search term contain special words or solely numbers that would not help in the search.

The number four in the strategy was chosen to the maximum word count in a term that may refer to a location name after observing typical toponym patterns around the world. It is assumed that the occurrence of toponyms with more than four words tend to be rare.

One auxiliary database for keeping previously searched terms and avoiding duplicated requests was developed. Such database is quite useful during the query execution stage. The historical database stores the searched terms and also the timestamp and information about retrieved features from the VGI source for each searched term. As the VGI changes all the time, an expiration time for each searched term should be established. Once the expiration time is reached for a term, a new search in the external

VGI source would be performed in order to update the toponyms related to such search term. Hence, new searches are performed periodically in order to store updated toponyms into the gazetteer, keeping a continuous enrichment process.

The OSM Nominatim[61] service, which indexes the entire OSM spatial database and provides toponym lookup, was used as a VGI data source for the GeoSEn gazetteer enrichment. An API using Java programming language was implemented for automatically performing searches on such service and receiving the JSON responses with the geographical location features. The OSM was chosen instead of LinkedGeoData (LGD) because preliminary experiments revealed that the endpoint was not working and consequently no data could be searched.

Geographical data retrieved during the query execution stage is analyzed before being finally stored into the GeoSEn gazetteer. This analysis is performed during the retrieved VGI analysis stage and consists of discovering the level of the toponym in the new GeoTree and its direct ancestor into the GeoSEn gazetteer. Figure 4.18 shows the process flow of the retrieved VGI analysis in details. The analysis of a retrieved VGI feature consists of four steps: 1) the checking for duplicates; 2) the classification; 3) the definition of a direct ancestor; and 4) the search of candidates for children. The first step consists of looking for similar features into the GeoSEn gazetteer comparing its geometries and/or names with the retrieved feature. Duplicated features are discarded and not stored into the gazetteer.

The classification of a retrieved VGI feature involves the extended version of GeoTree structure as such feature needs to fit with one of the levels in the tree. For example, a retrieved feature can be classified as either a POI or a City. The classification step is based on both the VGI metadata classifier and the spatial-search-driven classifier.

The OSM metadata structure, for instance, provides some fields that can help to identify where a retrieved feature can fit in the GeoTree. An OSM feature that satisfies the condition "`class:place && type:hamlet`" in its metadata fields, for example, can be mapped into the District level in the new GeoTree. Another feature that has the field "`class:highway`", for example, can be mapped into the Street level.

[61] http://nominatim.openstreetmap.org/

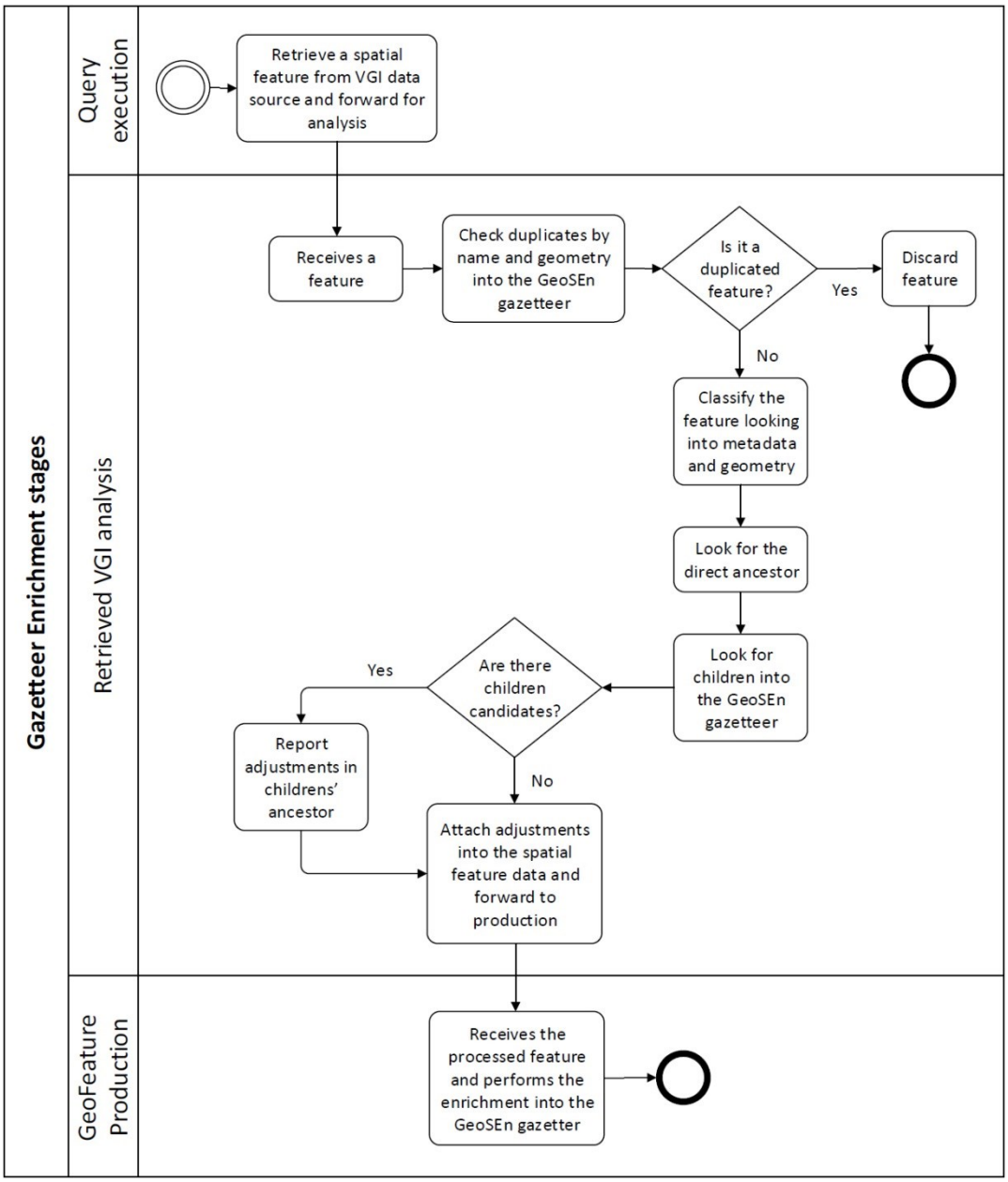


Figure 4.18: The process flow of retrieved VGI analysis stage

As OSM allows users to leave metadata fields blank or even to put wrong information, a classifier based on performing spatial searches within the GeoSEn gazetteer was developed. The main idea of this second classifier is to find out spatial relationships between new retrieved features and the previous stored ones by using known spatial operations such as contains, intersects, touches and overlaps. For instance, the geometry of a previously stored feature may overlap or contain the geometry for the new retrieved feature. Therefore, the spatial-search-driven classifier is applied in the step for definition of a direct ancestor of a new retrieved feature in order to search for a previously stored feature

suitable to be the direct ancestor of the retrieved feature in the GeoTree. The classifier finds the previously stored feature with the smallest GeoTree distance level above the level of the new feature, which spatially contains or overlaps such a new feature.

Finally, the search of candidates for children is applied to find previously stored features of which the new feature is an ancestor. This step is also based on the spatial-search-driven classifier. Obviously, this step does not apply to new features classified as POI because they could not assume the ancestor's role in the GeoTree. As the retrieved VGI analysis finishes, the last stage of the gazetteer enrichment method is responsible for generating the new toponym entry into the GeoSEn gazetteer and performing the updates on its children as well.

Once the GeoSEn gazetteer was enriched in order to store more precise toponyms in a city-context, it was necessary to perform some changes in the GeoSEn heuristics for toponym recognition and toponym resolution considering the extended GeoTree structure (shown in Figure 4.15).

The internal GeoTree production method, which is responsible for structuring the hierarchical tree of each toponym resolved, was remodeled. It included the three new toponym levels (POI, Street and District) so that the GeoTree production could recognize them correctly. The GeoSEn parser's module, which is responsible for checking cross-references between every toponym-resolved candidate, included the spatial distance calculation. Cross-references are geographic references found in a document that has topological spatial relationships in relation to other reference (Campelo and Baptista, 2009). The previous version of such module only considered the textual distance (word-by-word) and the hierarchical relationship between such candidates into a message. Hence, the spatial distance is used in addition to the textual distance only when two compared toponym candidates have a common ancestor and at least one of these candidates is classified as POI, Street or District, from the new toponym types added.

For example, given two tuples in the format ("*name*", "*type*") representing a toponym-resolved candidate: ("*Eiffel Tower*", "*POI*") and ("*Rue la Fayette*", "*Street*"). In such case, both candidates have at least one common ancestor in the GeoTree: *Paris* (of the city level). Thus, as the spatial distance decreases, the cross-reference coefficient for both candidates increases.

As described in Section 2.2 (Chapter 2), the GeoSEn system uses a metric called Confidence Rate (CR), which represents the probability of a toponym-resolved candidate to be a valid place. The CR value varies between 0 and 1, inclusive. In the GeoSEn system, each toponym level in the GeoTree has a specific CR calculation. A specific CR for each toponym type is necessary as there are several different ways that influence people mentioning a toponym depending on such type. Therefore, the CR calculation for each toponym-resolved candidate from each one of the new included toponym levels in the GeoTree was defined.

As pointed out in Section 2.2, the CR calculation is based on a set of Confidence Factors (CFs) and that each one has a weight in the CR computation for each toponym type. As the focus of this research is on the GeoSEn parser and not on the search engine provided by the GeoSEn system, the CF_{TS} was not considered on this study. It was needed to define the weights for CFs in the CR computation for the additional toponym types. As CR value may vary between 0 and 1, the weight sets were defined as shown in Equations 4.10, 4.11 and 4.12:

$$CR_{(POI)} = 0.40\, CF_{ST} + 0.15\, CF_{FMT} + 0.45\, CF_{CROSS} \quad (4.10)$$

$$CR_{(Street)} = 0.25\, CF_{ST} + 0.15\, CF_{FMT} + 0.60\, CF_{CROSS} \quad (4.11)$$

$$CR_{(District)} = 0.30\, CF_{ST} + 0.15\, CF_{FMT} + 0.55\, CF_{CROSS} \quad (4.12)$$

The weights of the previous toponym types found in the GeoSEn system were considered on defining such weight sets. The addition of the spatial distance for CF_{CROSS} calculation involving the additional toponym types, as well as the empirical nature of occurrence of these types in text messages were also considered.

Another change performed was implementing a cut function in order to discard less-geographically-precise toponym-resolved candidates while geoparsing microtexts. Although such redundancy is relevant for the search engine of the GeoSEn system, it is unsuitable in the approach to urban issues identification, which requires high geographical precision for toponym resolution. The main idea here is to keep only the most precise toponym-resolved candidates. For example, given the tweet "`very beautiful view`

`from Eiffel Tower in Paris`", there will exist two toponym-resolved candidates ("*The Eiffel Tower, Paris, France*", "*POI*") and ("*Paris, France*", "*City*"), and both candidates will be resolved at the end of the geoparsing process. Once the GeoTree knows that the Eiffel Tower is in Paris (France), it is not necessary to resolve for the less precise toponym even they are both present in the source text.

Finally, some changes were performed in order to enable the GeoSEn parser to deal with English documents. A set of special terms and another set of stop words from English language were introduced in the GeoSEn heuristics. The language of the document to be parsed is then an additional input for the parser.

The Geographical Analysis Process

In addition to the locations mentioned in the message, a tweet may also be attached to a location regarding the place where the user posted the message (the geocoded location), and a location obtained from the user profile (the user home location). While the tweet location is generally provided in terms of geographical coordinates (latitude and longitude), the user home location is usually freely provided by the user and generally requires being geoparsed. The analysis of such location contexts enables to identify toponym candidates to be resolved for the geographical facet of a tweet in the AGI production. Figure 4.19 illustrates how the geographical analysis proceeds.

The geographical analysis first looks into the tweet metadata in order to identify geocoded locations. Such metadata analysis enabled to find out that current tweets can be attached with two kinds of geocoded locations: geographical coordinates (precise) and geo entities (imprecise). While the geographical coordinates is geographically precise as it points directly to a point on the map, the geo entities comes from a Twitter gazetteer and they are geographically imprecise as it presents bounding box areas instead of points. Such bounding boxes generally relate to the city or country boundaries, which are imprecise to be used in the urban toponyms context.

The geographical analysis prioritizes precise geocoded locations for assigning the geographical facet. Geoparsing is only used when the tweet does not provide a precise geocoded location, as it can be noticed in the process flow shown in Figure 4.19. Such analysis design for the geographical facet was defined due to the results found in an

analysis on a labeled sample of tweets (see Section 5.3 in Chapter 5) that has investigated the relationship among tweet location contexts.

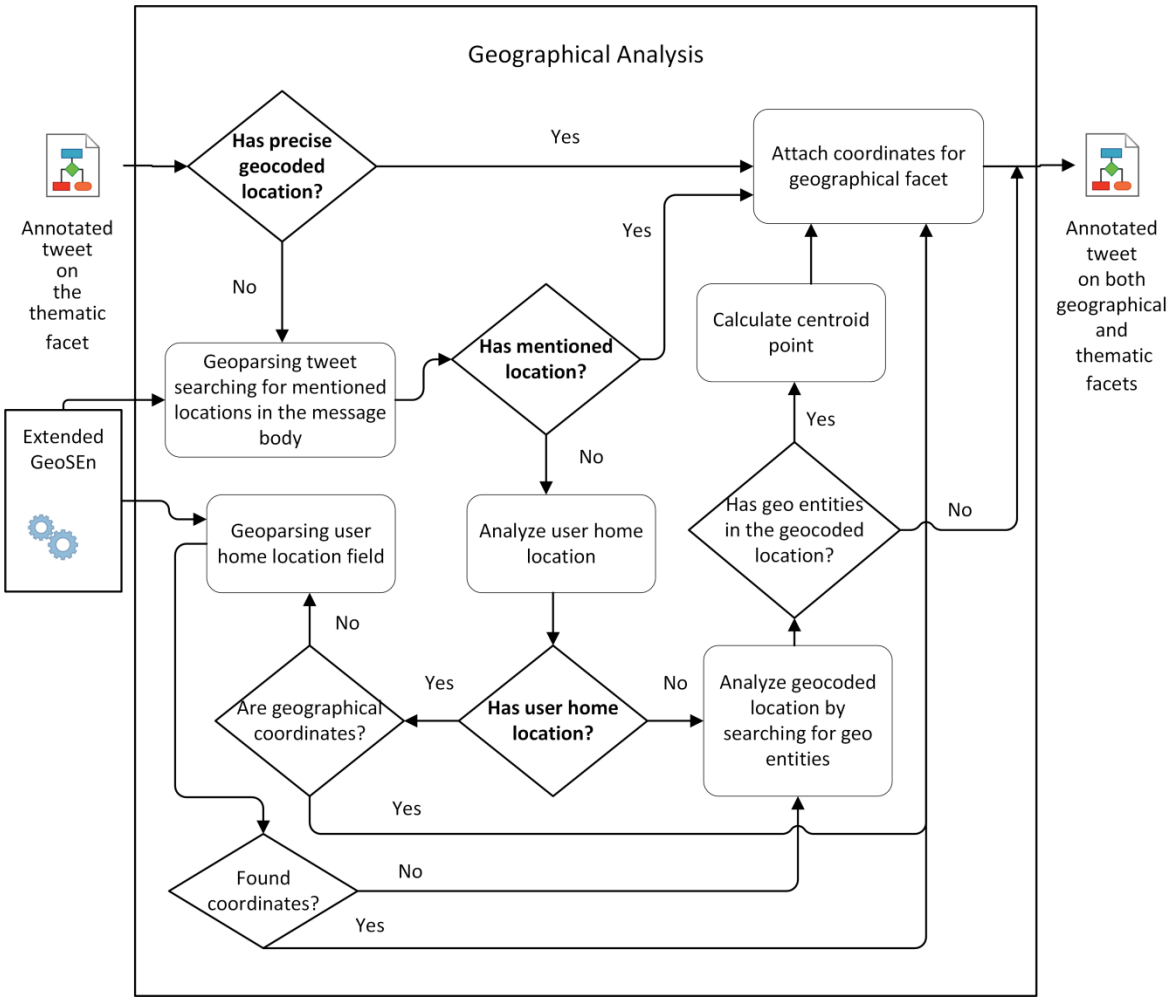


Figure 4.19: The geographical analysis process flow

The developed extended GeoSEn system is adopted to perform geoparsing in two cases: on searching for mentioned locations in the tweet message body, when there is no precise geocoded location; and on searching toponyms from the user home location field, when the user did not filled up such field with geographical coordinates. On both cases, the extended GeoSEn tries to resolve toponyms at a high GLoD, suitable for urban areas.

Listing 4.5 presents a JSON output example from the geographical analysis in a tweet without precise geocoded location that has a mentioned location found in the tweet message body. The JSON output produced from this task is merged with the JSON produced in the thematic analysis task.

Listing 4.5: Example of JSON output produced by the geographical analysis task

```
{
  "tweet_id": 651873880924143617,
  "toponyms": [
    {
      "point": [
          53.324238, -6.3857873
      ],
      "srid": "EPSG:4326",
      "keywords": [
        {
          "word": "Dublin",
          "position": "9"
        }
      ],
      "geoparsed": true,
      "confidence": 0.85,
      "id": 158155,
      "name": "Dublin, County Dublin, Leinster, Ireland",
      "geotree_level": "3",
      "parent_id": "234512"
    }
  ]
}
```

In the sample JSON (Listing 4.5), the attribute `toponyms` is an array with all the locations resolved from a tweet. The attribute `id` refers to the place name identifier in the gazetteer. The attribute `name` is the place name found in the gazetteer. The attribute `parent_id` refers to another place name in the gazetteer which has a parent relationship according to the GeoTree. The attribute `keywords` contains a list of words extracted from the tweet that enabled the geoparser to produce such a candidate. The attribute `point` contains the geographical coordinates for the toponym as a 2D point. In cases where the place name geometry is a point, such point is directly used to fill up this attribute. Otherwise, the centroid point is calculated from the place name geometry or bounding box and then used. The attribute `geoparsed` is a Boolean attribute that indicates whether the resolved toponym is a geoparsing result (true) or geographical coordinates that come directly from tweet metadata (false).

Finally, the attribute `srid` indicates the Spatial Reference System (SRS) related to the geographical coordinates presented in the attribute `point`. The default `srid` is the

EPSG:4326, which is related to the WGS84[62] geodetic system adopted by both Twitter and GeoSEn parser.

The attribute `confidence` refers to the confidence level returned by the geoparser ranging [0,1]. In case of geoparsing user home location, such confidence value is the half from the confidence level returned by the geoparser. Such assumption was designed in order to decrease the confidence while geoparsing user home locations. A study carried out on labeled tweet datasets showed that home user location is not reliable in comparison with the other tweet location contexts.

When a tweet contains precise geocoded location, only one toponym is returned in the output JSON with solely `point` and `srid` attributes filled up, and with the attribute `geoparsed` set to `false`. The other attributes are set to `NULL`. On the other hand, when a tweet does not contain precise geocoded location, all the locations found by the geoparser on searching mentioned locations or looking at user home location are sent to the summarization stage. Furthermore, a tweet that does not present any location candidates has the `toponyms` attribute equals to `NULL`. Those tweets are discarded and the other tasks are aborted for such tweet.

4.3.5 Temporal Analysis

The temporal analysis consists of exploring the temporal facet of the urban issues for the AGI production from tweets. As the temporal facet is not part of the main scope of this research due to time constraints, the many aspects and challenges regarding the temporal facet are not explored in this work.

The temporal analysis aims at providing accurate temporal marks for the urban issues identified in the thematic analysis task. In a similar way to the geographical facet, the timestamp of the social media post may not match the time of occurrence of an urban issue report. Thus, instead of just using the tweet timestamp, this task relies on a temporal tagger in order to identify time expressions eventually mentioned in the tweet message bodies. Such expressions may provide more accurate time ranges in comparison with the tweet timestamps.

[62] http://spatialreference.org/ref/epsg/wgs-84/

The Stanford Temporal Tagger was adopted though the SUTime library (Chang and Manning, 2012) for recognizing and normalizing time expressions. SUTime was chosen due to the support provided for English expressions and the Java compatibility. For example, a tweet posted on "2016-10-25T16:29" with the text fragment in the message body: "next Friday at 10:30pm", can be resulted in the timestamp "2016-10-28T22:30".

The tweet timestamp is given to the SUTime as the current reference time. Therefore, mentioned expressions such as "*last week*", "*for two weeks*", "*yesterday*" among others, could be parsed into a timestamp to be attached in the AGI production. As Twitter works with the Greenwich Mean Time (UTC+0), the produced timestamps are from such time zone. Listing 4.6 shows an example of a JSON output produced by the temporal analysis task. Similar to the geographical analysis, the JSON output produced from this task is merged with the JSON produced in the analysis task previously performed.

Listing 4.6: Example of JSON output produced by the temporal analysis task

```
{
  "tweet_id": 651873880924143617,
  "temporal": {
      "initial_time": "2016-10-28T22:30:00+0",
      "final_time": "2016-10-28T22:30:00+0"
  }
}
```

It can be noticed that the temporal analysis task always returns a time interval. Both initial and final times are the same when the interval is solely a timestamp. In order to simplify the processing of the temporal analysis task, only one time interval is returned. On tweets which contain more than one time expressions, the closest parsed timestamp to the tweet timestamp is chosen. Finally, the temporal analysis uses the tweet timestamp as output for tweets which does not contain time expressions.

4.3.6 Summarization

The summarization task is achieved only in case a tweet contains at least an urban issue candidate identified during the thematic analysis task and also if this tweet contains at least a location candidate identified during the geographical analysis task. The summarization task tries to resolve eventual semantic conflicts regarding both thematic and

geographical facets of urban issues reported in tweets. Figure 4.20 presents the process flow for the summarization task.

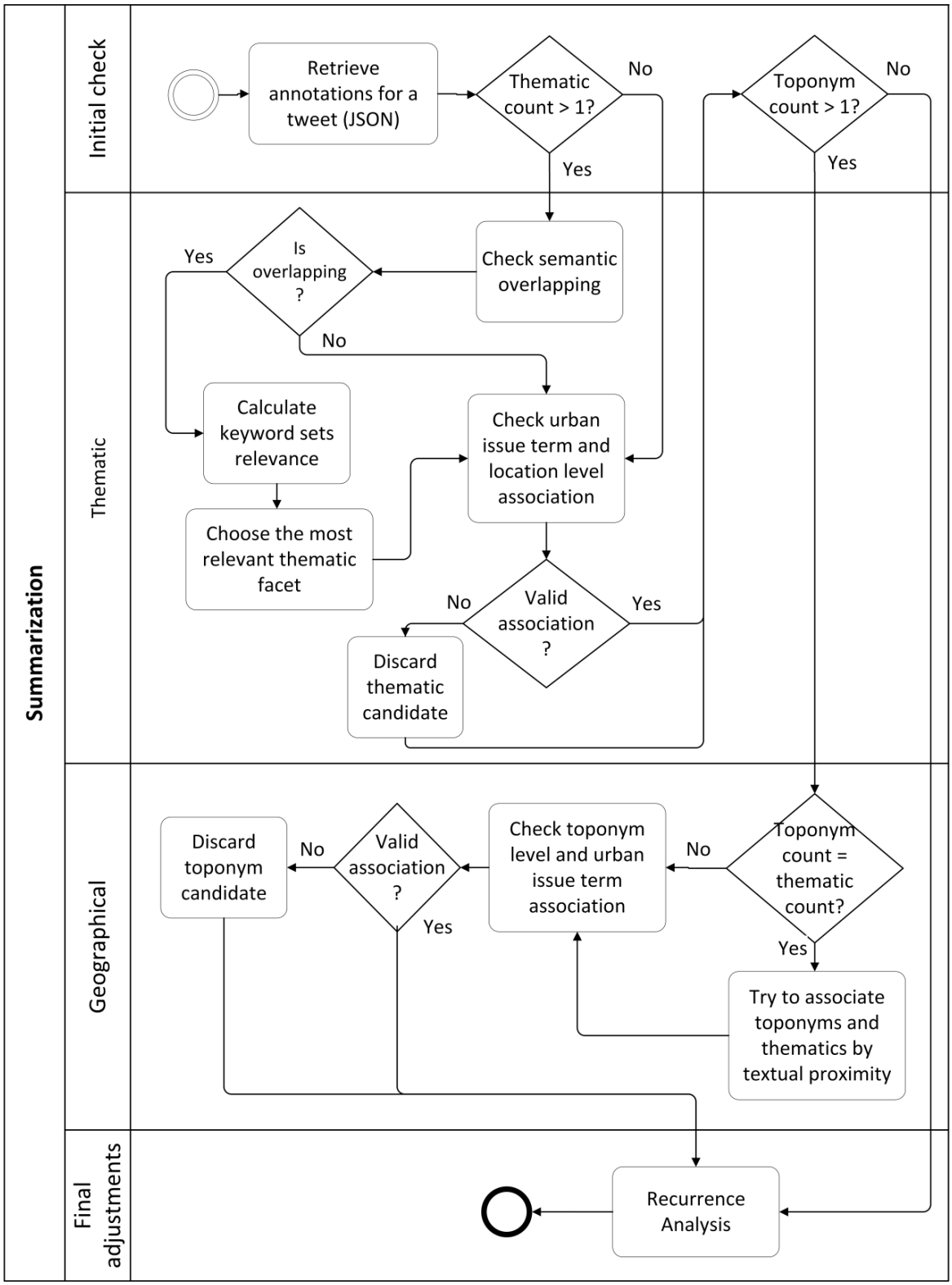


Figure 4.20: The process flow for the summarization task

The summarization task can be split into two main subtasks: resolving thematic and geographical conflicts. Thematic conflicts may exist when there are two or more different urban issue types identified for the same tweet. For example, the thematic analysis can identify a "path defect" reports and also a "litter/rubbish" report. This thematic conflict needs to be analyzed in order to discover whether there are two or more urban issues reported at the same tweet or whether the detected issues are overlapping each other in the tweet.

In order to analyze and identify thematic conflicts, the urban issue candidates are sent to checking semantic overlapping. Such checking consists of verifying whether there is semantic overlapping between words in the keyword sets from detected urban issue types. For example, the thematic analysis on a tweet with a fragment "*broken lamp in the pavement*", may result in two urban issue reports: "*broken lamp*" (a street lighting issue) and "*broken pavement*" (a path defect issue); a case of semantic overlapping, since the same token appears in the keyword sets from both urban issue reports returned. Therefore, in case of non-overlapping, the summarization considers the tweet is reporting different urban issues and consequently different AGI to be produced. In case of semantic overlapping, the urban issue candidates are forwarded to the relevance selection.

The relevance selection consists of calculating the relevance of the keywords set for each urban issue candidate and then selecting the candidate that presents the highest relevance in the tweet. The keyword set relevance is calculated by the function defined in Equation 4.13:

$$Relevance_{(issue_{type},tweet)} = \frac{\frac{|K|}{|W| - |S|}}{maxDistance(K)} \quad (4.13)$$

Given:

K: the keyword set from the urban issue candidate;

W: the words set from the tweet;

S: the stopwords set from the tweet;

$maxDistance(K)$: a function which returns the maximum occurrence distance between the elements of K in the tweet.

For example, the tweet "*there is a broken lamp in the pavement*" ($W = 8$ and $S = 5$) has two urban issue reports found in the thematic analysis: "*broken lamp*" ($K = 2$) (a street lighting issue) and "*broken pavement*" ($K = 2$) (a path defect issue). Although the K value is the same for both urban issue reports found, the $maxDistance(K)$ values are 1 and 3, respectively. Thus, the calculated relevancies are 0.66 for the street lighting issue and 0.22 for the path defect issue.

The summarization of the thematic facet also checks the association defined by the Equation 4.5 in order to identify whether the urban issue candidates are valid according to the toponym candidates found. The UIDO ontology provides the needed knowledge for such association through the relation ***occursIn***`(UrbanIssueTerm, UrbanLocationLevel)` (See Figure 4.7). Therefore, all the urban issue candidates which do not have such relation satisfied are discarded. If no urban issue candidate remains, the tweet is also discarded and no AGI is produced. Otherwise, the summarization proceeds to check if there are geographical conflicts.

Geographical conflicts may indicate that the tweet mentions or the user home location refers to two or more different geographical locations. In this case, there is no possibility of place name overlapping because the geoparser is able to solve it in advance during the geographical analysis.

Two or more different toponym candidates may indicate that the Twitter user is reporting the same urban issue type at different locations or that the tweet contains two or more urban issue reports, each of which associated with a distinct geographical location. Thus, if the number of toponym candidates is the same as the number of urban issue candidates, the summarization first tries to associate each toponym with a different urban issue according to their textual proximity in the tweet. Otherwise, in case that there are two or more urban issue reports and just one toponym, such a toponym is assigned to all urban issues found. Finally, in case that there is just one urban issue report and two or more different toponyms, all of them are considered as the Twitter user may be reporting in a single tweet that the same issue affects multiple places.

The association between toponym candidates and urban issues is tested using the knowledge for the UIDO ontology. In this particular case, toponym candidates which do

not match any urban issues resolved are discarded. Otherwise, the valid toponym candidates together with the urban issues are sent to the recurrence analysis.

The recurrence analysis consists of searching into the AGI produced database for similar urban issue at the same location in a different time span. As stated in Equation 4.2, the temporal nearness is one of the mandatory factors to group urban issue reports and to classify them as the same urban issue, according to a given time span. Such time span must be satisfied among the urban issue reports when they are about the same urban issue. Otherwise, there is a case of a recurrent urban issue.

It is hard to define the suitable time span between urban issue reports for the same urban issue. The time span between the grouped urban issue reports can be one day, one week or even one month. As the temporal facet is minimally addressed in this research, a further work should focus on defining such time span.

Finally, `recurrence = 0` means the urban issue is punctual in a specific location. On the other hand, the reported urban issue may be getting chronic as the number of recurrence increases. The number of recurrence is then attached to the urban issue candidates, which are finally forwarded to the AGI production.

4.3.7 AGI Production

The AGI production stage is responsible for creating the output of the proposed approach: urban issue reports from analyzed social media messages (tweets). Such a report consists of a spatiotemporal marker that can be read and analyzed by many digital systems such as LBSNs in the smart cities domain. Listing 4.7 presents an example of such output in JSON format.

As shown in Listing 4.7, the output is formed by the original tweet message body harvested from the social network, the urban issue information, the urban issue location information, and the valid time inferred for the urban issue report.

The AGI producer finishes the process of automated identification of urban issues from social media illustrated in Figure 4.13 and detailed along this chapter. Appendix C details a system prototype (Social2AGI) implemented to verify the applicability of the proposed approach.

Listing 4.7: Example of JSON output produced by the proposed approach

```
{
  "tweet_id": 651873880924143617,
  "tweet_text": "Huge potable water leaking near 142 O'Connell
                 Street Upper Dublin",
  "urban_issue_type": "leaks/drainage",
  "recurrence": 0,
  "location": {
      "name": "Dublin, County Dublin, Leinster, Ireland",
      "point": [ 53.324238, -6.3857873 ],
      "srid": "EPSG:4326",
      "confidence": 0.85
  },
  "valid_time": {
      "initial": "2016-10-28T22:30:00+0",
      "final": "2016-10-28T22:30:00+0"
  }
}
```

4.4 Summary

This chapter presented the proposed solution for the problem of automated urban issues identification from social media data. The proposed approach addresses current issues regarding social media data processing, such as preprocessing techniques, thematic, geographical and temporal analysis, followed by a summarization that resolves ambiguity issues in the knowledge discovering.

The knowledge discovered from the proposed approach is called AGI because the geographic information from social media posts combined with the semantics of urban issues tends to be AGI, a deviation from VGI, as stated by Stefanidis et al. (2013). Therefore, social media users assume the role of volunteers in the production of AGI, which could automatically be made available to users of LBSN applications. The proposal was generalized for AGI production since the thematic analysis can be applied in other domains by changing the domain ontology.

The urban issues domain was defined formally and the Urban Issues Domain Ontology (UIDO) was proposed and developed in order to enable urban issues identification from social media through a thematic parser. The next chapter discusses the evaluation of this research and presents the results acquired from some case studies.

Chapter 5

Evaluation

This chapter presents the evaluation of this thesis proposal through case studies. It is structured as follows: Section 5.1 presents the datasets used in this research; section 5.2 presents the evaluation of the thematic facet of the proposed approach to identifying urban issues from social media; Section 5.3 presents the evaluation of the geographical facet concerning the usage of the extended GeoSEn system and providing a discussion about the geographical contexts that can be related to a tweet; finally, Section 5.4 presents a summary of this chapter.

5.1 Datasets

Aiming at making real data available for the evaluation of the proposed approach to the automated identification of urban issues from social media, a crawler system was developed in Java language in order to harvest FixMyStreet and Twitter data. Thus, this research focused on crowdsourced data in English and in the geographical context of two English-speaking cities: Dublin, in the Republic of Ireland; and London, in the United Kingdom. The English language was chosen because this is the most spoken language worldwide, whilst the geographical contexts were chosen due to:

1) a one-year scholarship held in Dublin, which allowed a better understanding of the region's geography, as well as the possibility of recruiting native volunteers;
2) Dublin and London have FixMyStreet instances continuously used by the population; and
3) London is the 7th city in the global ranking of cities by the total percentage of all georeferenced tweets and the 2nd city worldwide when considering only cities that use English as an official language (Leetaru et al., 2013).

The details regarding each dataset are explained in the following.

5.1.1 FixMyStreet

In order to collect urban issues data for the UIDO ontology learning process, the FixMyStreet dataset was acquired from the LBSN instances running in Greater Dublin (IE) and Greater London (UK). Unfortunately, FixMyStreet only provides public access to the last 20 reports for a specific urban area through RSS feeds. Faced with this issue, a crawler system was developed to collect the FixMyStreet reports, which kept collecting information from the LBSN and storing the reports in a local PostgreSQL RDBMS. Listing 5.1 presents an example of a FixMyStreet report extracted from a RSS file harvested.

Listing 5.1: An example of a FixMyStreet report from Greater Dublin

```
<item>
   <title>Street Lamp, 19th June</title>
   <link>http://fixmystreet.ie/report/5653</link>
   <description>
        Street Light outside number 17 Valley Park Drive is not
        working
   </description>
   <category>Street Lighting</category>
   <pubDate>Fri, 19 Jun 2015 15:07:47 +0000</pubDate>
   <georss:point>53.381066238623 -6.3166263478135</georss:point>
</item>
```

As illustrated by the example shown in Listing 5.1, each FixMyStreet report is composed of six attributes:

- **Title:** a free text defined by the user submitting the complaint, followed by the publishing date;
- **Link:** the full URL for the report shared in FixMyStreet;
- **Description:** a free text description where the user describes the complaint into details;
- **Category:** the urban issue type defined by the user through a pre-defined list shown and defined by each one of the FixMyStreet instances;
- **PubDate:** the full timestamp when the report was created;
- **Georss:Point:** the geographical footprint of the complaint. This spatial information is provided by the complainant user through mouse clicks on an interactive map.

In large cities, FixMyStreet provides specific RSS feeds of reports for each city region. Thus, Dublin City metropolitan area is compound by four different RSS feeds, each one for a city council. The London metropolitan area is compound by 34 different RSS feeds, each one for a city borough. The list containing all the URLs of the RSS feeds used by the crawler is available in Appendix D.

The crawler ran continuously during the entire year of 2015. Each one of the 38 related RSS feeds was checked for updates every 30 minutes. Such 30-min time frame size was defined based on preliminary tests on several different sizes (10-min, 15-min, 1-hour and 6-hours), which helped to find the best choice that could avoid losing reports (more than 20 new reports between the time frames). Thus, for each RSS feed checking, the crawler read each one of the 20 reports, stored the new ones and discarded those ones that were already stored in other time. Table 5.1 describes the two datasets of FixMyStreet reports.

As it can be seen on Table 5.1, the crawler collected 18,130 reports from FixMyStreet during 2015. There were too many categories in the FixMyStreet instance running in the UK whilst the Republic of Ireland's instance had only six. Considering that FixMyStreet report categories are modeled as urban issue types in the UIDO ontology (see Chapter 4), a semantic mapping among the categories from both urban areas explored was established.

Table 5.1: FixMyStreet reports harvested for this research

Urban Area	Time Period	RSS Feeds	Total reports	Distinct categories
Greater Dublin (IE)	Jan/2015 to Dec/2015	4	1,564	6
Greater London (UK)	Jan/2015 to Dec/2015	34	16,566	65
		38	18,130	71

An analysis conducted with the 70 original categories of the dataset from Greater London suggested that some categories could be merged. For example, 5 distinct categories found in the reports ("Street light", "Street lights", "Street lighting", "Street Lighting" and "Faulty street Light") could be grouped into a single issue type "Street Lighting".

Other similar cases were also found from the report dataset. Such redundant categories were merged as part of a data cleaning process performed manually, which considered the semantics of each category. Such a cleaning process also enabled to discover categories from Greater London reports that group many different categories into a unique issue type. For example, the category "Public toilets" groups reports that fit better in other categories such as "Leaks and Drainage" or "Rubbish/Litter".

Therefore, five categories ("Parks repairs", "Parks/landscapes", "Public toilets", "Zebra crossing" and "Other") and their respective reports were removed from the dataset since it was challenging to reorganize them into the correct categories for two reasons. First, there were 727 reports to be analyzed manually; second, the semantics found on the reports was mixed (many reports from categories with distinct semantics mapped into a single one), making it difficult to classify without domain specialists. Therefore, the original amount of reports harvested from Greater London (17,293) was reduced to 16,566, as stated in Table 5.1, whereas the number of distinct categories was reduced from 70 to 65.

On the other hand, the six distinct categories from Greater Dublin dataset ("Graffiti", "Leaks and Drainage", "Litter and Illegal Dumping", "Road or Path defects", "Street Lighting", and "Tree and Grass maintenance") are quite consistent with their semantics. Hence, a manual mapping between the grouped categories from the Greater

London and the distinct categories from the Greater Dublin datasets was performed. For example, the categories "Fly Posting", "Flyposting", "Graffiti" and "Graffiti and flyposting" from Greater London reports were mapped to the category "Graffiti" from Greater Dublin reports. Such category mapping enabled to merge the datasets according to their urban issue types. Table 5.2 shows the urban issue types which comprise the FixMyStreet report dataset.

As it can be observed in Table 5.2, the most expressive issue types are related to litter and road defects, since both issue types represent together 77.5% of the dataset.

Table 5.2: FixMyStreet report dataset grouped by urban issue types.

Urban Issue Type	Reports	Average report length (words)	Kind of urban issues covered
Graffiti	1,189	20	Graffiti, fly posting and other types of visual pollution in unpermitted places
Leaks and Drainage	259	36	Any kind of leaks and blocked or faulty drains or gullies, generally on streets or roads
Litter and Illegal Dumping	8,878	26	Accumulated litter, dumped rubbish, dog fouling, fly tipping, gritting, overflowing litter bin and missed bin collection, generally on streets, roads, pavements, and open properties
Road or Path defects	5,173	41	Potholes, manholes, faulty road/street signs, street furniture, abandoned vehicles, illegal parking, bollards, bus stop problems, slippery tracks, obstructions, defects on pavements/footpaths and traffic lights
Street Lighting	1,477	28	Broken street lamps and faulty street lights
Tree and Grass maintenance	1,154	44	Overhanging foliage/vegetation, public trees/grass needing pruning, problems with street weeds and floral displays

5.1.2 Tweets

In order to acquire tweets that may refer to urban issues and then evaluate the UIDO ontology, a crawler was developed using the Twitter Streaming Java API[63], aiming at harvesting geocoded tweets using the bounding box-driven query provided. Unfortunately, the current geographical query supported by the Twitter API only works with bounding boxes instead of precisely-shaped polygons.

The adopted method for data collection differs from many proposals in the state-of-the art as it collected a small percentage of the Twitter stream that are directly geo-referenced into specific coordinates (geocoded). The idea behind collecting solely geocoded tweets is to provide a corpus from homogeneous urban areas, language, and culture. Thus, tweets from Greater Dublin (IE) and Greater London (UK) were harvested. The crawler ran continuously during the entire year of 2015, in parallel with collecting FixMyStreet reports, crawling tweets every minute for each selected urban area separately.

Each tweet was extracted from the JSON format to a record in a relational table stored in a PostgreSQL RDBMS. Since the original amount of harvested tweets (around 35 million) is too large to be manually classified by humans, a sample dataset was extracted from those harvested tweets. The filtering and manual labeling strategies applied on both tweet datasets in order to classify them into urban issues and mentioned geographical locations are described in details in Appendix A. A full sample of a tweet in JSON format from the acquired datasets is presented in Appendix B. Table 5.3 describes the two tweet datasets manually labeled regarding urban issues and geographical locations mentioned in the body. Table 5.4 shows the urban issue types which comprise the labeled tweet datasets.

Table 5.3: Manually labeled tweet datasets used in this research

Urban Area	Time Period	Total Tweets	Issue Types
Greater Dublin (IE) [-6.46,53.22,-6.03,53.45]	Dec/2014 to May/2015	403	6
Greater London (UK) [-0.51,51.31,0.30,51.71]	Jan/2015 to Dec/2015	1,091	6
		1,494	6

[63] https://dev.twitter.com/overview/api/

Table 5.4: The manually labeled tweets dataset grouped by urban issue types

Urban Issue Type	Tweets	Average tweet length (words)
Graffiti	89	18
Leaks and Drainage	106	20
Litter and Illegal Dumping	41	19
Road or Path defects	126	20
Street Lighting	129	20
Tree and Grass maintenance	6	20
(Not an Urban Issue)	997	20

The kind of issues covered in each urban issue type in Table 5.4 is equivalent to the one presented in Table 5.2. From Table 5.4, it can be noticed that 20 words is the average length of a tweet from the dataset. Considering the constraint of 144 characters per tweet, it means that each word is formed by around 7 characters in average. Finally, Table 5.5 presents the amount of tweets manually labeled to a geographical location, distributed per level of the extended GeoTree (presented in Chapter 4).

Table 5.5: Amount of tweets manually classified to a geographical location, distributed per the extended GeoTree levels

Extended GeoTree level	Tweets
Level 0 – Country	25
Level 1 – Province	0
Level 2 – County	12
Level 3 – City/Town/Village	125
Level 4 –District/Suburb	69
Level 5 – Street/Road	200
Level 6 – Point of Interest	199
(*Not a geographical location*)	864

From Table 5.5, it can be observed that most of the tweets that have mentions to geographical locations in the message bodies are related to locations inside urban areas (level greater than 3), for a total of 468 tweets, which represent 31% of the dataset and 74.3% of the tweets with geographical mentions. Figure 5.1 shows a Venn diagram that illustrates graphically the tweet gold standard dataset according to the contexts assigned by the human volunteers in the labeling process.

In Figure 5.1, the gray area represents tweets with no context assigned by the volunteers (no urban issue neither mentioned geographical location). The light green area represents tweets with only thematic context assigned (urban issues). The pink area represents tweets with only geographic context assigned (mentioned locations). Finally, the light green x pink intersection area represents tweets with both thematic and geographic contexts assigned, which means urban issue reports that have the location of the issue specified by the complainer.

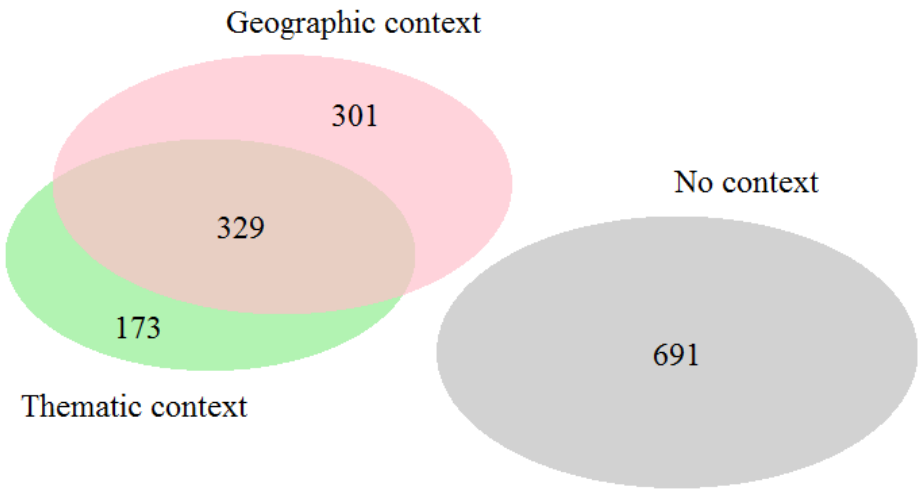


Figure 5.1: Venn diagram illustrating the gold standard dataset according to the contexts assigned

5.2 Evaluating the Thematic Facet

In order to evaluate the thematic facet of the approach to automated identification of urban issues, an evaluation of the UIDO ontology being used in the thematic parser was performed.

There are many ontology evaluation approaches. Haghighi et al. (2013) provide a critical review of those approaches, which can be categorized into six main classes: gold standard-based; data driven; evaluation by humans; application-based; task-based; and criteria-based. Haghighi et al. (2013) also discuss the need for choosing the suitable approach according to the domain and the main purposes of the developed ontology, or even to consider a derived approach, since no single one might perfectly fit all the objectives of ontology evaluation. Therefore, the evaluation of the UIDO ontology followed the application-based approach, since the main aim of the proposed domain ontology is to provide automated urban issue identification in texts.

The UIDO performance on identifying urban issues in the developed thematic parser was compared with the following machine learning classifiers widely used in the literature: Naïve Bayes (John and Langley, 1995), Multinomial Naïve Bayes (MNB) (Mccallum and Nigam, 1998), Random Forests (Breiman, 2001) and Support Vector Machines (SVM) (Cortes and Vapnik, 1995). The k-Nearest Neighbors (KNN) was not used in this evaluation because it was outperformed by SVM in the work carried out by Jin et al. (2013) about customer complaints. Other mentioned machine learning classifiers widely used in the literature such as J48 and Bayesian Network were not used in such comparison due to time constraints in this research.

The datasets presented in Section 5.1 were used in a quantitative evaluation. The main purpose of the quantitative evaluation is to demonstrate the validity of the modeled terms and relationships. Thus, the statistically-based approach proposed for automated identification of relevant terms in the domain of urban issues could also be evaluated. For such, three evaluation setups were defined by using both FixMyStreet reports and tweets. Table 5.6 presents the key information in each evaluation setup.

Each setup consists of a *k*-fold cross-validation (Kohavi, 1995) with three repetitions. In a cross-validation, the datasets are split into *k* mutually exclusive sub datasets (folds) with the same size and proportion of urban issue types or urban issues. The data that compose each sub dataset is selected randomly in each repetition. The number of *k* in each setup was different due to the amount of data used in each setup and to the need of splitting the datasets into equal amounts. As there are 497 tweets labeled as urban issues, a 7-fold schema produces sub datasets with 71 urban issue tweets in each one.

Table 5.6: Evaluation setups for the thematic facet of the proposed approach

Setup	Training dataset	Test dataset	Classes	Folds (*k*)	Repetitions
1	FixMyStreet reports	FixMyStreet reports	6 classes (urban issue types)	10	3
2	Tweets	Tweets	2 classes (yes/no urban issues)	7	3
3	FixMyStreet reports	Tweets	2 classes (yes/no urban issues)	7	3

Fourth and fifth setups using six (the issue types) and seven classes (the issue types and yes/no urban issue) were initially planned but was discarded because of the low amount of tweets grouped per issue types, as it can be seen in Table 5.4. The "Tree and Grass maintenance" issue type, for instance, groups solely six tweets.

The hardware used to run these setups was a desktop computer from the Information Systems Laboratory (LSI/UFCG) with an Intel Core i7-3770 3.9 GHz Processor, 32 GB DDR-3 DRAM and 3 TB HDD running Windows 7 64 Bits. The software used to run the setups was developed in Java language and includes the Weka (Hall et al., 2009) Java API and the LibSVM (Chang and Lin, 2011) for Java. In order to run the evaluation setups with Naïve Bayes, MNB and Random Forests, the algorithms from the Weka Java API were adopted and ran using their default input parameters. In order to run the evaluation setups with SVM, the LibSVM for Java was adopted. The input parameters for the SVM are detailed in each evaluation setup.

Aiming at providing a fair comparison among the performance of the five classifiers involved in this study, the data preprocessing techniques used in both the UIDO ontology learning process and the thematic parser were adopted to preprocess the input data for the machine learning classifiers. Moreover, the Bag-of-Words (BoW) representation, a simplest and almost universally used approach (Boulis and Ostendof, 2005), was adopted to create the vector data embedded in the arff (supported by Weka) and the libsvm input file formats.

In the BoW representation, each message (tweet/report) is represented by a vector $M = (w_{1m}, w_{2m}, w_{3m}, \ldots, w_{vm})$. The v is the size of the vocabulary (distinct word count in

each dataset), and the w_{im} is the weight of a word i in a message m. Such word weight is calculated by the traditional *tf-idf* statistic (Spärck Jones, 1972):

$$w_{im} = TF_{im} \times IDF_i = \left(1 + \log_2(f_{im})\right) \times \log_2(\frac{N}{n_i}) \quad (5.1)$$

In Equation 5.1, f_{im} is the frequency of a word i in a message m, N is the total number of messages (tweets/reports) in the dataset, and n_i is the number of messages that contains the word i.

Finally, four statistical measures for each setup were calculated: Accuracy, Precision, Recall, and F-score (or F-measure, or F1-score); widely used to evaluate classification techniques (Baeza-Yates and Ribeiro-Neto, 1999). Those metrics are calculated based on the number of positives and negatives in a classification task. Starting from a class of interest (urban issue type or yes/no urban issue), True Positives (TP) are tweets/reports from this specific class that are correctly classified in this specific class. True Negatives (TN) are tweets/reports from other classes that are correctly classified in these other classes. False Positives (FP) are tweets/reports from other classes that are wrongly classified in the class of interest. False Negatives (FN) are tweets/reports from the class of interest that are wrongly classified in other classes. In the following, the mathematical formulas for the first three metrics are described in details (Baeza-Yates and Ribeiro-Neto, 1999):

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \quad (5.2)$$

$$Precision = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Positive} \quad (5.3)$$

$$Recall = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Negative} \quad (5.4)$$

The precision and recall are the bases for the F-Score calculation. The F-Score is the harmonic mean of precision and recall and thus summarizes the performance of the classifier in a [0,1] value range. The F-Score formula is given by:

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5.5)$$

Therefore, the F-Score is the metric adopted to evaluate and compare the classification strategies in each one of the three developed evaluation setups, since it better represents the performance of the classifiers. The F-Score was computed for each fold in each repetition. The following subsections explain in details each evaluation setup and present the comparative results among the five classification strategies used. This section ends with a discussion regarding the results from the evaluation of the thematic facet.

5.2.1 Setup 1: FixMyStreet multi classification

A cross-validation was carried out using the widely adopted *k*-fold schema with 10 folds. Unfortunately, it could not be possible to split the 18,130 FixMyStreet reports into 10 subsets (folds) with the same amount of reports (1,813). As it can be noticed from Table 5.2, the quantities of reports per issue type are not divisible per 10.

In order to maintain the same proportion of reports per issue type in each fold, the FixMyStreet dataset had to be adjusted by performing undersampling. The following reports were randomly crossed out in setup 1: 9 reports from "Graffiti" urban issue type; 9 reports from "Leaks and Drainage"; 8 reports from "Litter and Illegal Dumping"; 3 reports from "Road or Path defects"; 7 reports from "Street Lighting"; and 4 reports from "Tree and Grass maintenance" urban issue type. At total, 18,090 FixMyStreet balanced reports remained for the evaluation.

In each one of the 3 repetitions performed, the FixMyStreet dataset was split into 10 mutually exclusive folds containing 1,809 reports each, randomly selected based on the issue type. As a result, 30 subsets were created and stored in distinct tables in a PostgreSQL RDBMS with a standard name "`fixmystreet_10folds_fold_`*`n`*`_rep_`*`r`*". The similarity between the same folds in different repetitions was around 20%. The training and test datasets were then produced according to these subsets and following the cross-validation *k*-fold schema (e.g. in fold 1, the training dataset is composed by the 9 other folds and the test dataset is composed by the fold 1). Thus, each training dataset contained 16,281 reports while the test dataset contained 1,809 mutually exclusive reports.

The arff files, the input data format required for the Weka implementation of Naïve Bayes, MNB, and Random Forests classifiers, were produced by using a script developed in Java. Such script read the datasets from the RDBMS, performs the same preprocessing strategy used for the thematic parser, translates the reports to the BoW representation and put them into vectors of *tf-idf* statistics. Each report was translated into a vector of 15,175 attributes filled up by real numbers in addition to the class, an integer number in the range [0,5], where each integer represents an urban issue type. Such arff files were also converted to libsvm format using a script in Python in order to produce the suitable input for the LibSVM implementation of the SVM classifier. At total, 60 arff files and 60 libsvm files were produced.

In order to run the thematic parser powered by the UIDO ontology, each training dataset were used to generate a version of the UIDO ontology with the learning process described in Chapter 4. Therefore, 30 owl files were produced. Each produced owl file has around 527 elements (words), which composed 445 urban issue terms. For the evaluation of the owl files, the thematic parser read the test datasets directly from the RDBMS. This fact prevented the need for producing 30 test datasets.

The arff files were enough to run Naïve Bayes, MNB, and Random Forests classifiers using their default input parameters, as well as the owl files were enough to run the thematic parser powered by the UIDO ontology. For the SVM, it was still needed to estimate the best cost (*-c*) and gamma (*-g*) parameters prior to run the C-SVC svm type with radial basis function as kernel type (*default*). To this aim, the Grid Search in each one of the training datasets was executed. Interestingly, the estimated best cost and gamma were the same in each one of the 30 training datasets: $c = 8.0$ and $g = 0.125$. It means that the 11.1% of distinction in each training dataset was not enough to change these SVM parameters.

Figure 5.2 presents the boxplots that show the variation of the F-Score per classifier and issue types. As it can be observed, the classifier which most varied during the training-test executions was the Random Forests. Therefore, Random Forests is the least reliable in Setup 1, followed by the MNB. On the other hand, it seems that SVM and the UIDO-driven classifiers where the best ones in Setup 1. SVM overcame the UIDO ontology in four of the six issue types.

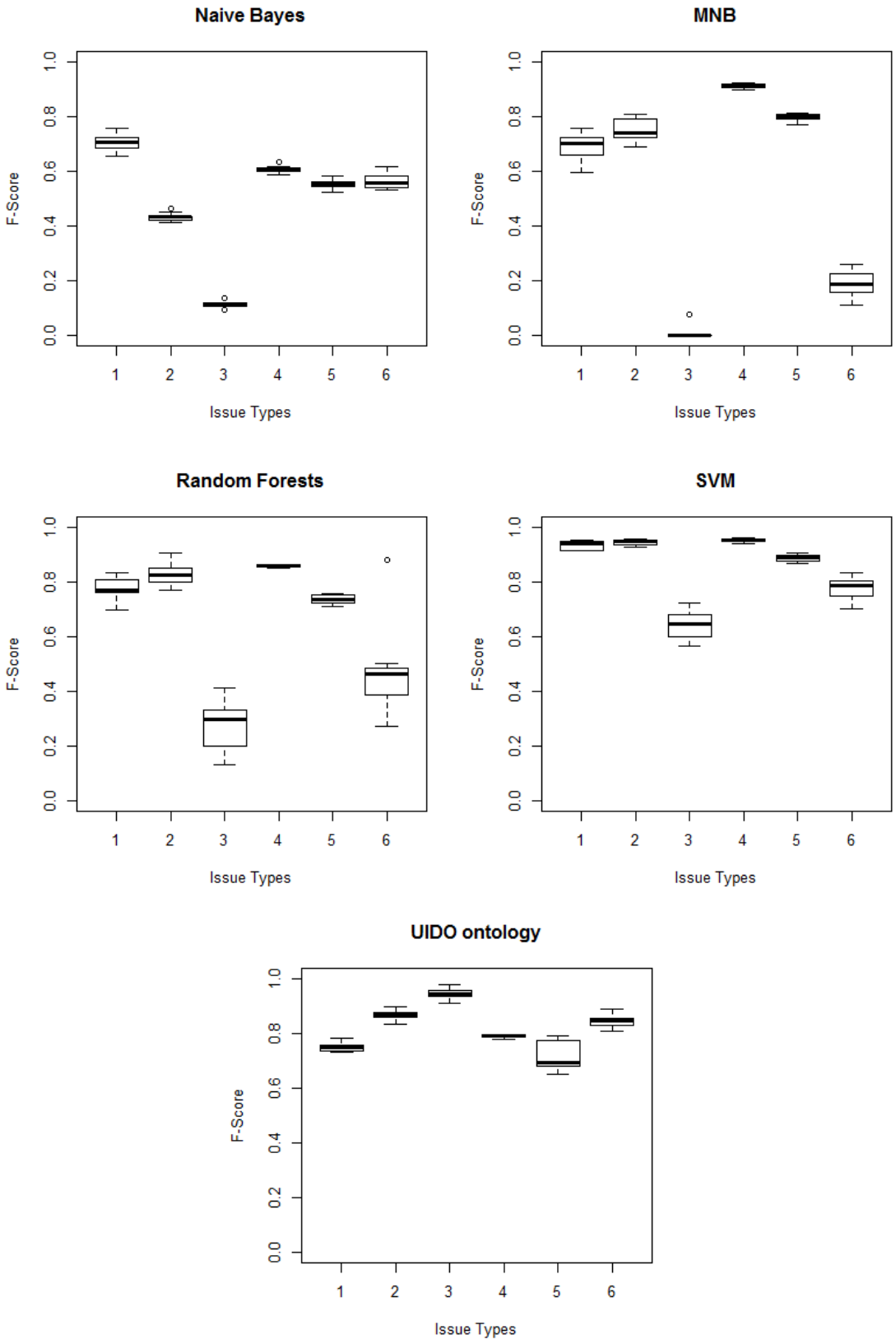


Figure 5.2: Boxplots per classifier and issue types: 1) Street Lighting 2) Graffiti 3) Leaks and Drainage 4) Litter and Illegal Dumping 5) Road or Path defects and 6) Tree and Grass maintenance

The boxplots shown in Figure 5.2 also highlight an issue regarding urban issue type 3 (Leaks and Drainage). The four machine learning classifiers presented lower F-Scores for such urban issue type classification. The MNB was the worst, with most of the F-Scores ranging close to 0. However, it could be still noticed that the UIDO-driven classifier presented high F-Scores for such issue type. The fact which can explain such observation is the number of reports for Leaks and Drainage (259, as it can be seen in Table 5.2). Such a number of reports is quite lower than other issue types and this fact may affect the machine learning classifiers, which work better with balanced classes. This problem has not occurred in the ontology learning, since the learning strategy focuses on identifying relevant terms in each issue type separately.

Figure 5.3 shows a line chart with the average F-Scores of the evaluated classifiers in each one of the 10 folds. Such chart shows that the SVM was the best classifier in Setup 1, followed by the UIDO-driven classifier, which presented similar results, both above 0.8. The other three classifiers presented F-Scores bellow 0.7, with Random Forests better than Naïve Bayes and MNB.

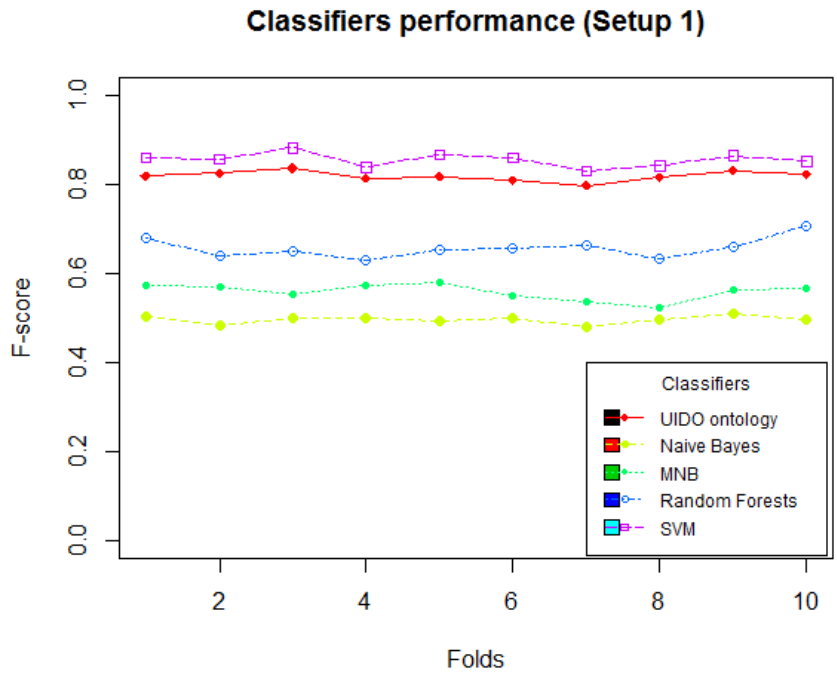


Figure 5.3: Line chart of classifiers performance on 10-folds cross-validation

In order to statistically verify whether the UIDO-driven and SVM classifiers could be considered distinct, the confidence intervals for the F-Scores with 95% of

confidence level were calculated. The shapiro-Wilk test (Shapiro and Wilk, 1965) was used to analyze whether the F-Score distributions for each one of the classifiers follows a normal distribution. A resulting *p-value* greater than 0.05 means the null hypothesis that the data distribution follows a normal distribution cannot be rejected. The analysis of normality in the F-Scores distributions was aided by histograms, density charts and normal Q-Q plots provided by R. Figure 5.4 shows the F-score confidence intervals calculated for each one of the classifiers.

Figure 5.4 aids to confirm that the SVM classifier overcame the UIDO-ontology driven classifier with the mean F-Score 0.0384 better. However, the confidence intervals from both classifiers intersect each other.

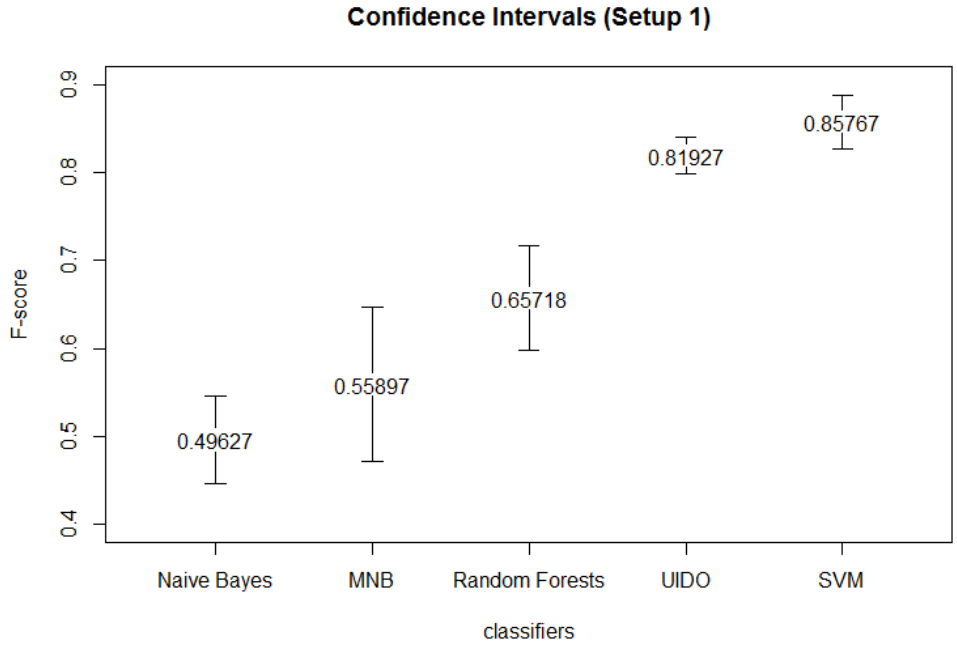


Figure 5.4: F-score confidence intervals for the each classifier in Setup 1 (95% confidence)

The Wilcoxon signed rank test (Wilcoxon, 1945) was then applied to statistically compare both classifiers. The resulting *p-value* of 0.104 (> 0.05) suggests that the null hypothesis which states that there is no difference between the classifiers cannot be rejected. Thus, there is a statistically trend of similarity between these classifiers.

The Wilcoxon signed rank test was still applied to compare the UIDO-driven classifier with Naïve Bayes, MNB and Random Forests. The resulting p-values were below 0.05, which rejected the null hypothesis.

A preliminary analysis over the thematic parser's log for setup 1 showed a number of false positive cases due to absent terms for specific urban issue types. For instance: "dumped car", which was classified as an urban issue from the "Litter and Illegal Dumping" type due to the term "dumped" and the mention to "bin" in the message.

There were also a number of false negatives due to absent terms, but less than false positives in general. Some adjustments in the building strategy for the UIDO ontology would reduce such a number of false positives and consequently increase the results. However, a deep analysis is needed to identify what need to be changed specifically.

5.2.2 Setup 2: Tweet binary classification

In setup 2, a cross-validation was carried out using the *k*-fold schema with 7 folds. Unfortunately, it was not be possible to split the 1,494 tweets into 10 subsets (folds) with the same amount likely in setup 1. As there were only 497 tweets manually labeled to urban issue types (see Table 5.4), k=7 was adopted as 497 is divisible per 7 and all the urban issues tweets could be used. Thus, each fold could contain 71 urban issue tweets to be tested by the same classifiers explored in setup 1.

Regarding the urban issue types, the amount of tweets per issue type was too small to be split into training and test datasets among the folds. The "Tree and Grass maintenance" issue type, the smallest one, could not even have one tweet per folds. Even the "Street Lighting" issue type, the biggest one, could have only 18 tweets per fold. Faced with this problem, a binary classification was carried out in setups 2 and 3, as they make use of the tweet dataset. In a binary classification, the classifiers need to check a tweet and classify as Yes/No regarding urban issue report features. Therefore, all the 497 manually labeled tweets to urban issues are from class 1 (Urban Issue: Yes), regardless the issue types assigned.

Considering that machine learning classifiers works well with balanced datasets, a binary classification with a balanced dataset was carried out. Undersampling was performed in order to make the tweet dataset balanced, which implies that some of the remaining 997 tweets manually labeled as non-urban issues had to be filtered out in order

to keep solely 497 tweets. These 497 were randomly selected and composed the dataset with class 0 (Urban Issue: No). Finally, the setup 2 used a dataset of 994 tweets. Such dataset was randomly split into 7 mutually exclusive folds containing 142 tweets, 71 from each class.

As three repetitions were performed in setup 2, 21 subsets were created and stored in distinct tables in a PostgreSQL RDBMS with a standard name "`tweet_7folds_fold_`*n*`_rep_`*r*". The similarity between the same folds in different repetitions was around 15%. The training and test datasets were then produced according to these subsets and following the cross-validation *k*-fold schema (e.g. in fold 1, the training dataset is composed by the 6 other folds and the test dataset is composed by the fold 1). Thus, each training dataset contained 852 tweets, while the test dataset contained 142 mutually exclusive tweets.

Similar to setup 1, the arff and libsvm files were produced in order to train the machine learning classifiers. At total, 42 arff files and 42 libsvm files were produced for training. Each tweet was translated in a vector of 3,287 attributes filled up by real numbers in addition to the class, an integer number in the binary range [0,1].

In order to run the thematic parser powered by the UIDO ontology, only tweets from class 1 (Urban Issue: Yes) from each training dataset were used to generate versions of the UIDO ontology. Thus, 21 owl files were produced. Each produced owl file has around 38 elements (words), which composed 41 urban issue terms. Then, the thematic parser and the machine learning classifiers ran similar to setup 1. For the SVM tunning, the Grid Search ran in each one of the training datasets. Unlike setup 1, the estimated best cost and gamma was distinct among the 21 training datasets. The c varied from 2.0 to 2048.0, while the g varied from 3.05×10^{-5} to 0.5.

Figure 5.5 presents the boxplots that show the variation of the F-score per classifier and class (Yes/No urban issues). It can be observed that Random Forests was again the classifier that most varied during the training-test executions. The F-Score for classifying urban issues varied around 0.15. Therefore, Random Forests is also the least reliable in Setup 2, followed by the Naïve Bayes, which presented an outlier.

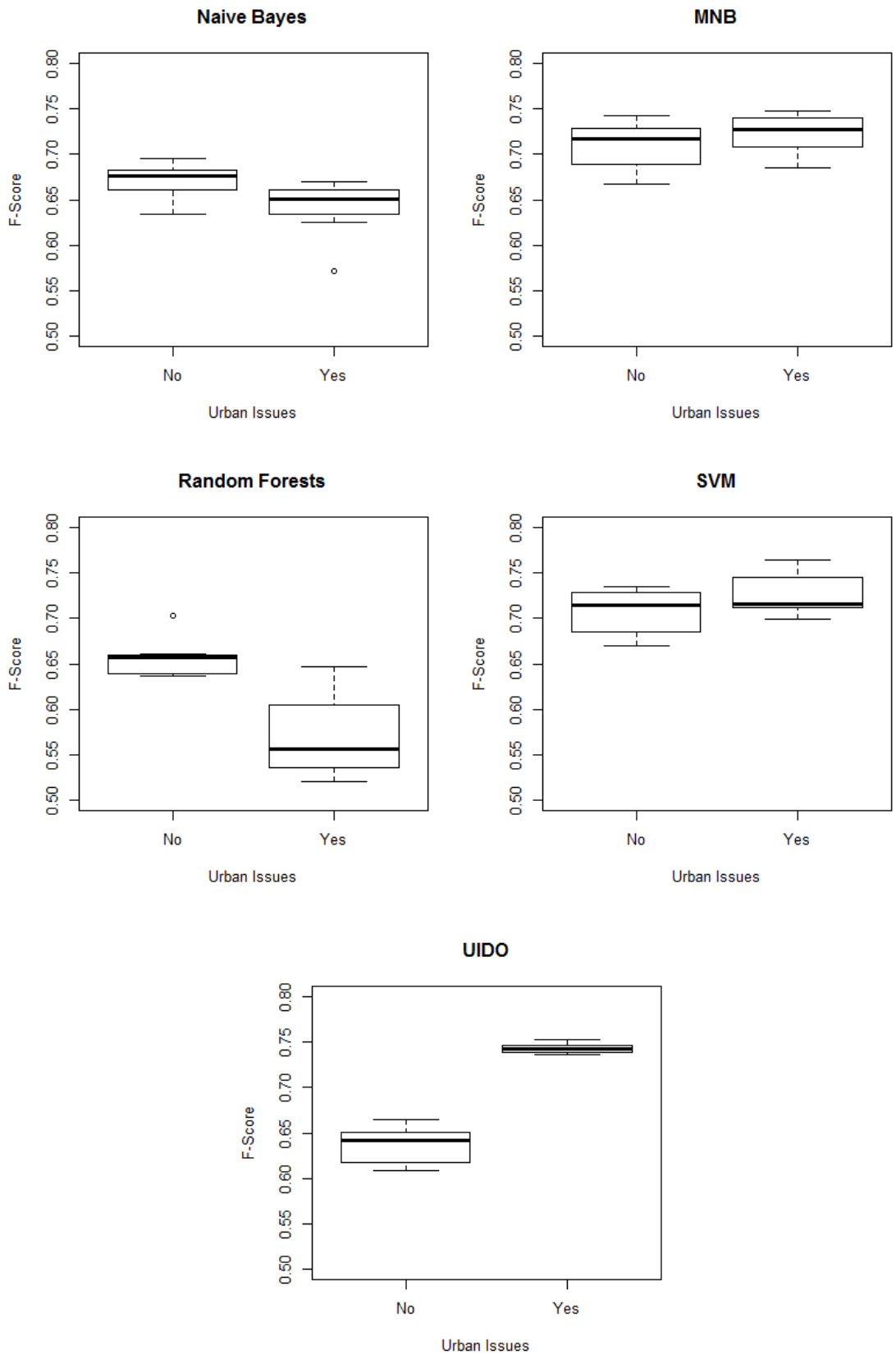


Figure 5.5: Boxplots per classifier and urban issues classification (Setup 2)

The other two machine learning classifiers (MNB and SVM) presented similar and balanced boxplots among Yes/No classes in Figure 5.5, with F-Score ranging at around 0.67 and 0.77. Finally, the UIDO-driven classifier presented a minimal variation on classifying urban issues, with F-Score ranging at around 0.74 and 0.76. However, the F-Score decreased by around 0.1 on classifying non-urban issues. This observation can be explained by the fact that the UIDO ontology learned solely using the urban issues dataset. Moreover, such observation also suggests there are an elevated number of false positives that means tweets classified as urban issues when they are not urban issues indeed.

This unbalancing observed in the classification of non-urban issues performed by the UIDO-driven classifier led it to be third algorithm according to the performance comparison illustrated by the line chart shown in Figure 5.6. Thus, the SVM and MNB were the best classifiers in setup 2, followed by the UIDO ontology.

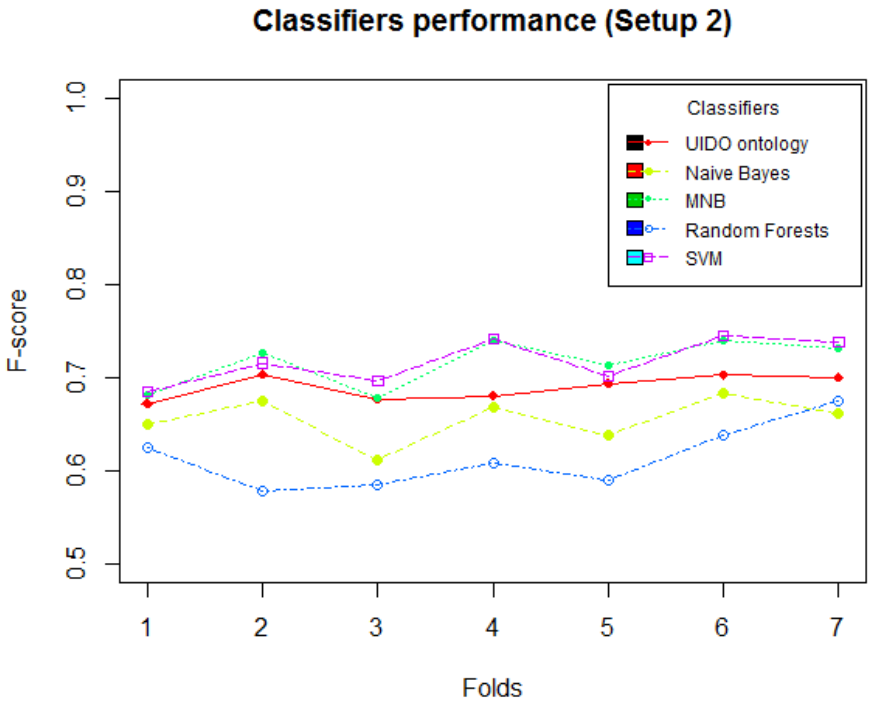


Figure 5.6: Line chart of classifiers performance on 7-folds cross-validation (Setup 2)

The line chart in Figure 5.6 shows the average F-Scores among the folds for SVM and MNB were very similar, while the UIDO was a bit worse although still near to SVM like setup 1. The other two classifiers were the worst in the performance comparison

through the average F-Scores, however, all of them presented F-Score above 0.57 among the 7 folds.

In order to statistically verify whether the UIDO-driven, the MNB and SVM classifiers could be considered distinct each other, the confidence intervals for the F-Scores with 95% of confidence level was calculated. Figure 5.7 shows the F-Score confidence intervals calculated for each one of the classifiers in setup 2.

The mean F-Score for SVM and MNB were very close, as expected after the boxplot and line chart analysis. The mean F-Score difference was of just 0.00177 and the confidence intervals from both classifiers not only intersect each other as they were quite similar. The UIDO-driven classifier was the third best algorithm, but close to the MNB and SVM as the F-Score differences were 0.02593 and 0.0277, respectively. However, the confidence intervals intersect each other among these three best algorithms.

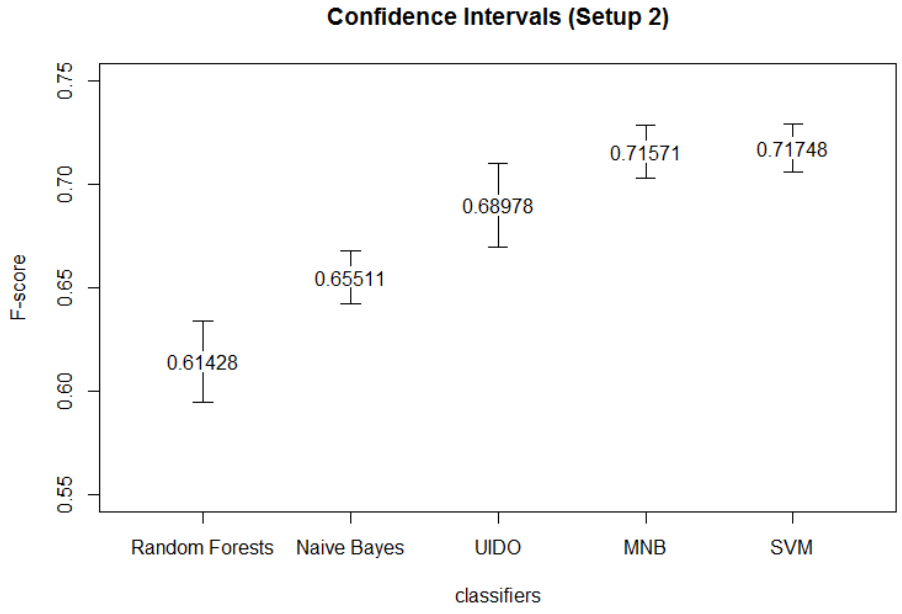


Figure 5.7: F-score confidence intervals for the each classifier in Setup 2 (95% confidence)

The Wilcoxon signed rank test (Wilcoxon, 1945) was then applied to statistically compare the three classifiers. The resulting *p-values* of 0.004895 (< 0.05) and 0.0203 (< 0.05) on comparing the UIDO-driven with SVM and MNB classifiers, respectively, suggest that the null hypothesis which states that there is no difference between the classifiers can be rejected. This implies that the SVM and MNB classifiers are statistically

different from the UIDO-driven. The Wilcoxon signed rank test was still applied to compare the UIDO-driven classifier with Naïve Bayes and Random Forests. The resulting *p-values* were also below 0.05 and rejected the null hypothesis.

A preliminary analysis over the thematic parser's log for setup 2 showed a number of false positive cases due to tweets with metaphor or sarcasm. Unfortunately, sarcasm is a challenge on text processing which has been addressed in many research proposals. Thus, a further work may include some metaphor/sarcasm processing method/tool from the literature in order to reduce false positive rates in this sense. The analysis also showed a number of false negative cases where the urban issue types were not already modeled by the UIDO ontology such as public services (e.g. mobile network outage) and traffic.

5.2.3 Setup 3: FixMyStreet and Tweet binary classification

The setup 3 intends to evaluate the classifiers on identifying urban issues reported in tweets but learning from FixMyStreet reports. Therefore, the both datasets used in setup 1 and setup 2 were used in setup 3.

Instead of performing a cross-validation using the *k*-fold schema with 10 folds like setup 1, only 7 folds were used due to the tweet dataset split issues discussed along the setup 2 description. As the FixMyStreet and tweet datasets are naturally mutually exclusive, the training and test subsets were established as the follows:

- Training: all the 18,130 FixMyStreet reports;
- Test: the 994 tweets split into 7 folds (the same test datasets from setup 2).

Regarding the urban issue types, all the FixMyStreet reports were labeled as urban issues though the issue types are from class 1 (Urban Issue: Yes). There was no corpus to be used in the training with elements from class 0 (Urban Issue: No), since all the FixMyStreet reports are urban issues being reported. One option for such elements would be using the 497 tweets with class 0 (Urban Issue: No) however, it would make the training dataset much unbalanced. Moreover, the training data in setup 3 is based in a one class classification. One class classification is a binary classification task for which only one class of objects, the target class or positive class, is available for training learning (Désir et

al., 2013). This scenario represents a disadvantage for machine learning classifiers, since only SVM can be setup to work with one class classification. On the other hand, the UIDO ontology does not face major issues since its learning process does not require labeled data.

Three repetitions were also performed in setup 3, although only the tweet test datasets were chosen randomly while the FixMyStreet training dataset kept constant. Thus, the 21 tweet subsets created in setup 2 were used for the tests. Similar to setup 1, the arff and libsvm files were produced in order to train the machine learning classifiers. At total, 22 arff files and 22 libsvm files were used in the training. Each report and each tweet was translated into a vector of 17,974 attributes filled up by real numbers in addition to the class, an integer number in the binary range [0,1].

In order to run the thematic parser powered by the UIDO ontology, all the FixMyStreet reports labeled as members of the class 1 (Urban Issue: Yes) were used to generate a version of the UIDO ontology. Therefore, only one owl file was produced from the learning with 18,130 reports. The produced owl file has 533 elements (words) which composed 451 urban issue terms. Then, the thematic parser and the machine learning classifiers except SVM ran similar to setup 1. The setup of the SVM changed to the One-Class SVM type, since there was only one class in the training data. The default SVM type (C-SVC) does not run with one class data. Besides the SVM type, it was still needed to estimate the best *nu* (*-n*) and *gamma* (*-g*) parameters prior to run the One-Class svm type with radial basis function as kernel type (*default*). The Grid Search algorithm then changed to search the best *nu* parameter instead of searching the best *cost* (performed in the previous setups). The best estimated *nu* and *gamma* for the training dataset were $n = 0.55$ and $g = 0.000122$.

Figure 5.8 presents the boxplots that show the variation of the F-score per classifier and class (Yes/No urban issues) in setup 3. It can be observed that MNB and Random Forests were very unbalanced, since the F-Scores on classifying tweets from the class 1 (Urban Issues: Yes) were around 0.7 while the F-Scores on classifying tweets from the class 0 (Urban Issues: No): were around 0.0. Such observation indicates that MNB and Random Forests classified all the tweets into the class they have learnt: “Yes”. Curiously, the One Class SVM also presented considerably unbalanced results, with most tweets classified into the class “No”.

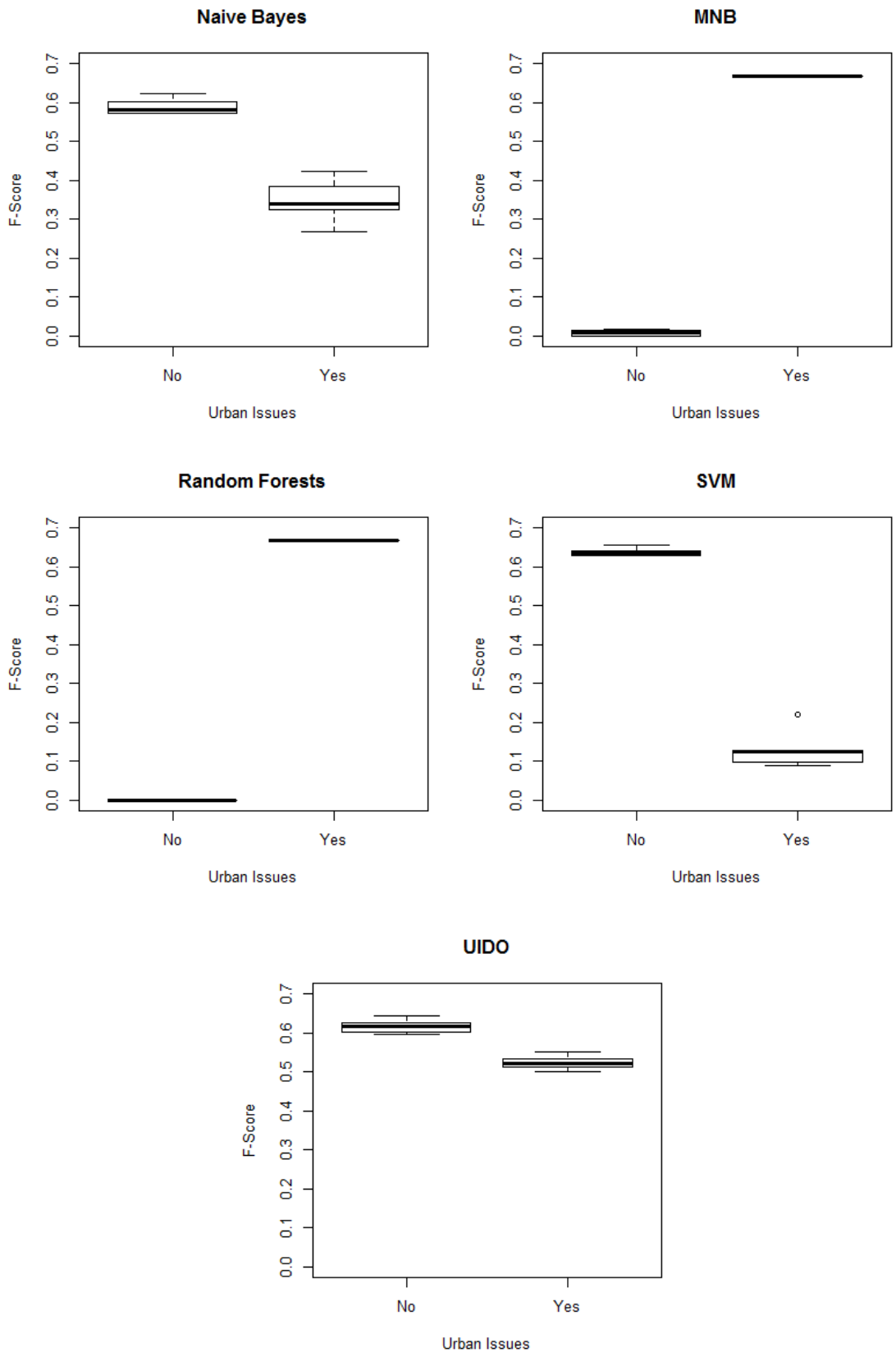


Figure 5.8: Boxplots per classifier and urban issues classification (Setup 3)

The UIDO-driven and the Naïve Bayes classifiers presented better boxplots, however, Naïve Bayes has presented a relevant unbalancing in the comparison among the F-Scores from the "Yes/No" classes. Moreover, Naïve Bayes already presented a notable F-Score variation for class "Yes", ranging at around 0.3 and 0.4. On the other hand, the F-Scores for UIDO-driven classifier presented low variation and unbalancing among the "Yes/No" classes. Curiously, the UIDO ontology performed better on classifying non-urban issues. Such observation may suggest an elevated FP rate on classifying urban issues.

Figure 5.9 shows a line chart with the average F-Scores from the classifiers along the 7 folds. It can be confirmed that the UIDO-driven classifier outperformed the other classifiers in the comparison. The MNB and Random Forests were the worst in setup 3, followed by the SVM. The line chart also makes clear the high variation of the average F-Scores presented by Naïve Bayes classifiers.

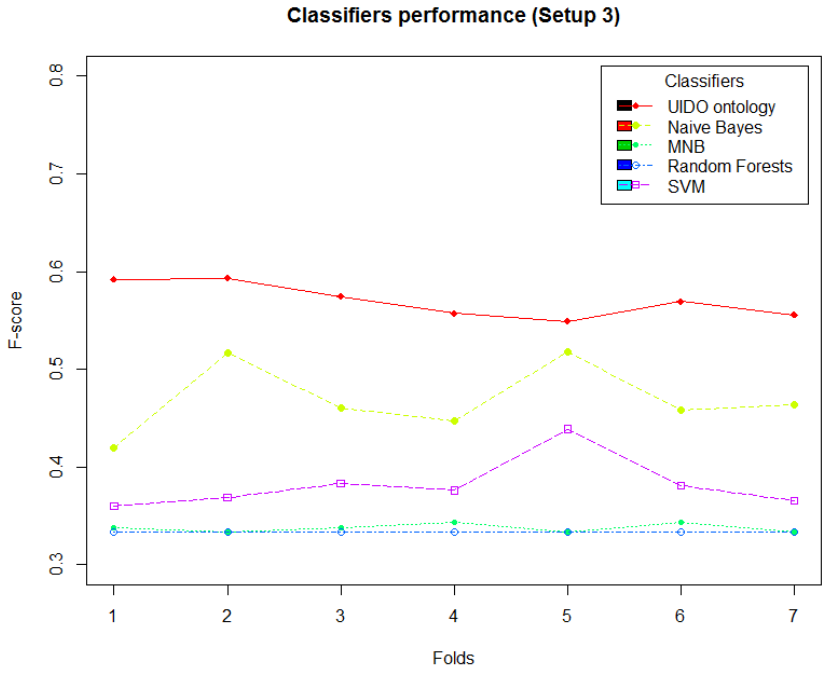


Figure 5.9: Line chart of classifiers performance on 7-folds cross-validation (Setup 3)

In order to statistically confirm whether the UIDO-driven classifier could be considered distinct from the other classifiers in setup 3, the confidence intervals were calculated for the F-Scores with 95% of confidence level. Figure 5.10 shows the F-score confidence intervals calculated for each one of the classifiers in setup 3.

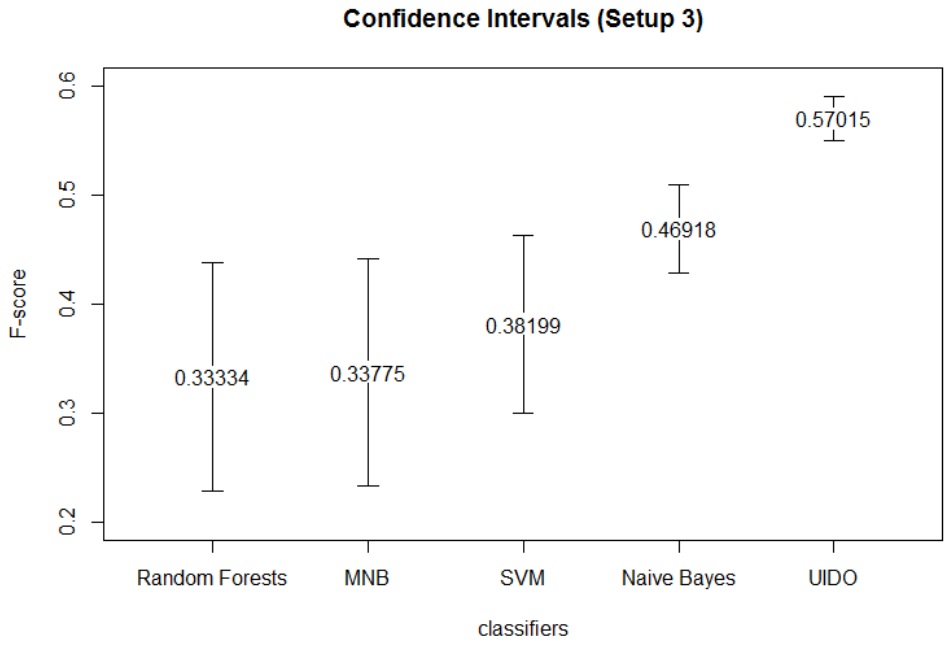


Figure 5.10: F-score confidence intervals for the each classifier in Setup 3 (95% confidence)

The mean F-Score for MNB and Random Forests were very close, as expected after the boxplot and line chart analysis. The mean F-Score for SVM was slightly better than the other two, but its confidence interval is almost totally contained in the intervals for MNB and Random Forests. Curiously, the confidence interval for Naïve Bays also intersects the intervals for these previous three classifiers. The confidence interval for the UIDO ontology is apart and small in comparison with the other ones.

The Wilcoxon signed rank test (Wilcoxon, 1945) was then applied to statistically compare the UIDO-driven with the other classifiers. The resulting *p-values* were all below 0.05, which suggests that the null hypothesis which states that there is no difference between the classifiers can be rejected.

A preliminary analysis over the thematic parser's log for setup 3 showed a number of false negative cases due to urban issue terms which were not modeled by the UIDO ontology. On most cases, there were synonym terms found on tweets which were not used in the FixMyStreet reports and consequently the ontology building process could not identify them. An evolving process may enrich the UIDO ontology with such a synonym terms from specialists in the domain or by performing the ontology building process with both tweets and FixMyStreet reports.

5.2.4 Discussion

The results obtained on running the three evaluation setups have shown that the performance of the UIDO-driven classifier was satisfactory in comparison with the Naïve Bayes, MNB, Random Forests and SVM classifiers. In Setup 1, which has considered a multi-class setting for classifying urban issues according to their urban issue types using the FixMyStreet dataset, the performance was competitive to the SVM classifier. With a mean F-Score above 0.8, the developed learning approach for the UIDO ontology from the FixMyStreet dataset proved satisfactory, since it was possible with around 500 attributes to obtain a similar performance of a classifier which relied on more than 15,000 attributes.

In Setup 2, which has considered a scenario using microtexts from Twitter for binary classification regarding the existence of urban issues being reported, the UIDO-driven classifier could not be as good as SVM and MNB, although it has presented a mean F-Score slightly below. The results have shown the need for further developments in the domain ontology and thematic parser in order to reduce the number of false positives.

In Setup 3, a variation of setup 2 that has considered the classifiers training using a larger and more reliable knowledge base, the UIDO-driven classifier performance overcame machine learning classifiers as expected. The scenario of setup 3 had not been good for the machine learning classifiers as only SVM supports one-class classification. However, the SVM classifier could not perform a good classification even using the one class setup.

Nevertheless, the scenario of setup 3 is the most adequate with real world, since the main intention on discovering urban issues from social media is to employ as most knowledge as possible for that particular domain. Thus, the knowledge learnt together from FixMyStreet, specialists and other LBSNs in the urban issues domain is crucial. This gives an advantage to the ontology-driven approach in comparison with the explored machine learning classifiers.

Finally, the mean F-Score at around 0.57 from the UIDO-driven classifier outperforms related work on concept finders such as 0.53 (Kang et al., 2014) and 0.28 (Cimiano and Volker, 2005).

5.3 Evaluating the Geographical Facet

This section describes the evaluation of the geographical facet in the proposed approach to automated identification of urban issues from social media. The evaluation of the geographical facet of urban issues focuses on the performance of the extended version of the GeoSEn system on geoparsing urban areas mentioned in tweets. Then, a complementary study is performed in order to identify the relationship among the three different geographical contexts that can be assigned to a tweet: the geocoded, the user home and the mentioned location. This section ends with a discussion regarding the results from the evaluation of the geographical facet.

5.3.1 The GeoSEn applied in tweets to identify urban area place names

The extended version of the GeoSEn system was tested on real case studies so that it could be evaluated on performing geoparsing of urban areas eventually mentioned in tweets. Two case studies were performed aiming at measuring the performance when geoparsing with more precise GLoD available in the improved gazetteer.

The first case study used the gold-standard tweet dataset from Greater Dublin (IE), while the second case study used the gold-standard labeled tweet dataset from greater London (UK). Thus, the 403 tweets from the Dublin dataset and the 1,091 tweets from the London dataset, containing 143 and 487 tweets manually assigned to a geographic location respectively, were processed by the extended version of the GeoSEn parser in order to automatically identify toponyms based solely on the body of tweets. Thus, 1,494 manually labeled tweets were geoparsed at total, 630 with mentions to toponyms.

The toponyms resolved automatically were compared with the data acquired from the manual labeling performed on each one tweet set. The true/false positives and negatives were computed in order to calculate the statistical measures regarding performance of the extended GeoSEn system. The confusion matrixes are provided in Table 5.7. True Positives (TP) means tweets that have geographical mentions and were geoparsed correctly. True Negatives (TN) means tweets that do not have geographical mentions and the geoparser correctly classified as without toponyms. False Positives (FP) means tweets that were incorrectly classified to toponyms. Finally, False Negatives (FN) means tweets that have geographical mentions but the geoparser incorrectly classified as without toponyms.

Table 5.7: Confusion matrixes for a) Dublin dataset and b) London dataset

True	False	
122	**55**	Positive
196	**30**	Negative

a)

True	False	
226	**114**	Positive
604	**147**	Negative

b)

Table 5.8 presents the four statistical metrics for geoparsing (Martins et al., 2005): Accuracy (Equation 5.2), Precision (Equation 5.3), Recall (Equation 5.4) and F-Score (Equation 5.5). These statistical metrics were calculated in order to measure the overall performance of the extended GeoSEn parser and its enriched gazetteer on English tweets.

Table 5.8: Statistical results for the case study on the extended GeoSEn system with English tweets

Dataset	Accuracy	Precision	Recall	F-Score
Dublin	78.9 %	68.9 %	80.3 %	0.742
London	76.1 %	66.5 %	60.6 %	0.634
(*weighted avg*)	77,1 %	67.4 %	67.9 %	0.674

It could be noticed that accuracy and precision were very near on both datasets. On the other hand, the recall has considerably decreased in the London dataset. Such finding can be explained by the high number of false negatives, which may indicate missing place names into the gazetteer. However, the weighted average F-Score at around 0.674 is a good result for the extended GeoSEn geoparser on geoparsing place names inside urban areas in comparison with related work. For instance, the F-Scores from TwitterNLP, Yahoo! PlaceMaker and Stanford NER 4-class on geoparsing tweets (Lingad et al., 2013) were 0.451, 0.540 and 0.576, respectively, without being enabled to identify place names in the urban GLoD.

5.3.2 Analyzing the relationship among tweet locations

A tweet may contain three different geographical contexts assigned: the geocoded, the user home and the mentioned location. The geocoded location is the location where the Twitter user posted the tweet and may be attached in the tweet metadata in form of

geographical coordinates. The user home location generally refers to the user hometown and can be filled up by the Twitter users in their profiles as strings. Finally, the mentioned location relates to place names eventually mentioned in the tweet. Several locations may be mentioned locations within a tweet, however, as the volunteers have labeled at most one in a tweet (see details in Appendix A), there is only one mentioned location related to a tweet in the labeled dataset.

While the geocoded location is provided in the form of geographical coordinates, the user home and the mentioned locations need to be geoparsed in order to be converted in geographical coordinates. As discussed in Chapter 3, geoparsing place names in urban area GLoD is challenging and the accuracies achieved by current geoparsers are not optimal. Therefore, this study intended to identify the relationship between the geocoded location and the mentioned location from a tweet.

The 630 tweets labeled with mentioned geographical locations were considered to carry out the study. Thus, all the tweets have geocoded and mentioned locations in form of geographical coordinates. As the user home locations were provided as strings, a manual geoparsing relying on Google Maps was performed in order to enable a complete analysis considering the three locations related to a tweet. Although it demands more manual work, the manual geoparsing was preferred instead of using the extended GeoSEn system because it is generally more accurate than automated methods. The manual geoparsing process included data cleaning, the usage of Google Maps Geocoding API[64] and disambiguation according to the human's inference on reading the tweets.

The geodetic distance provided by the `ST_DistanceSpheroid()` PostGIS function[65] was adopted for the spatial comparison among the locations considering the datum WGS 84[66] adopted by both Twitter and Google Maps. Given the approximate radius of Greater Dublin area from a point in the Dublin City Centre is 15 km, and the approximate radius of Greater London area from a point in the London City Centre is 30 km, six distance ranges were defined to normalize the calculated distances: less than 0.5km; between 0.5km and 1km; between 1km and 5km; between 5km and 15km; between 15km and 30 km; and more than 30km.

[64] https://developers.google.com/maps/documentation/geocoding/intro/
[65] http://postgis.net/docs/ST_Distance_Spheroid.html
[66] https://epsg.io/4326

Figure 5.11 presents box plot charts that show the distance distributions for each one of the three distances calculated in each tweet dataset: between the geocoded and user home locations; between the user home and mentioned locations; and between the geocoded and mentioned locations. It could be noticed that the distance between the geocoded and the mentioned locations are the shortest among the analyzed distances on both Dublin and London datasets. The median distances are 2.446 km and 3.596 km in Dublin and London datasets respectively. This means that the distance between the geocoded and the mentioned location in 50% of each analyzed dataset is up to around 2.5 km and 3.6 km. Moreover, the first quartiles for such distance are 0.441 km and 0.526 km in Dublin and London datasets respectively.

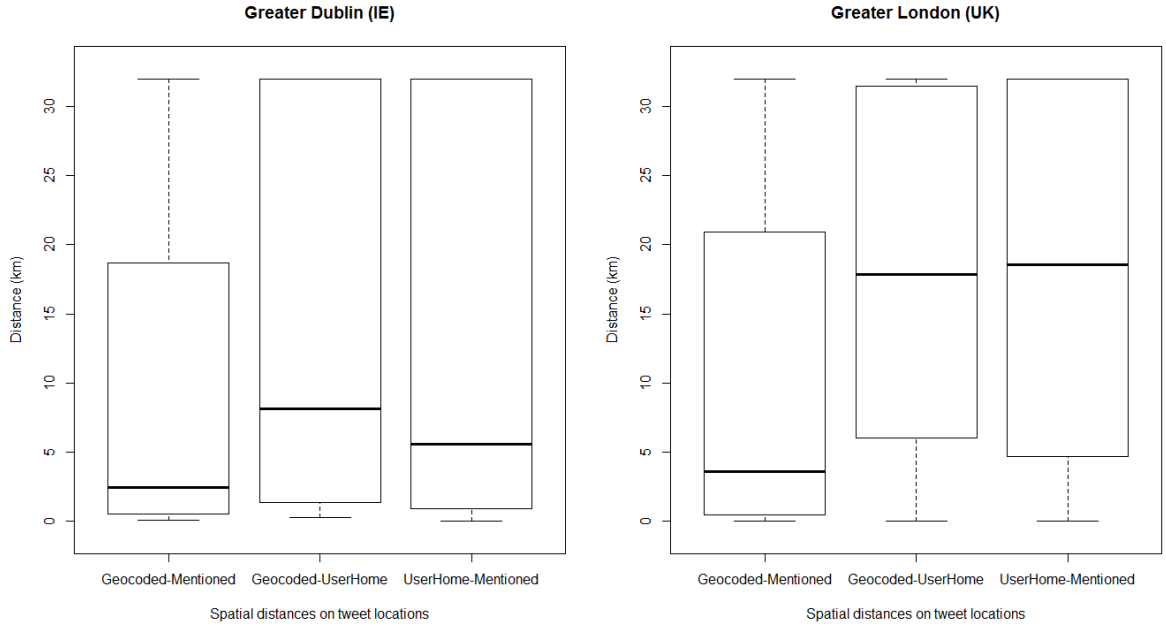


Figure 5.11: Boxplots of the spatial distance distributions in the tweet locations on both Dublin and London datasets

The user home location seems to be not reliable to be used in non-geocoded tweets when the context of urban areas is important since the median distances are over 5 km. However, the boxplots related to the user home locations shown in Figure 5.11 may indicate an interesting moving behavior of Twitter users on posting tweets in their cities. Regarding the comparison between the geocoded and the user home locations on the analyzed datasets, 50% of the Twitter users post their tweets in a distance within 8.1 km and 17.9 km from their homes in Dublin and London respectively. Regarding the comparison between the mentioned and the user home locations on the analyzed datasets,

50% of the Twitter users post tweets related to locations in a distance within 5.5 km and 18.6 km from their homes in Dublin and London respectively. Curiously, such distances are inside the respective urban areas.

Figures 5.12 and 5.13 shows comparative histograms that enable to visualize in more details the distribution for the three spatial distances in each one of the 6 distance ranges defined on both urban areas.

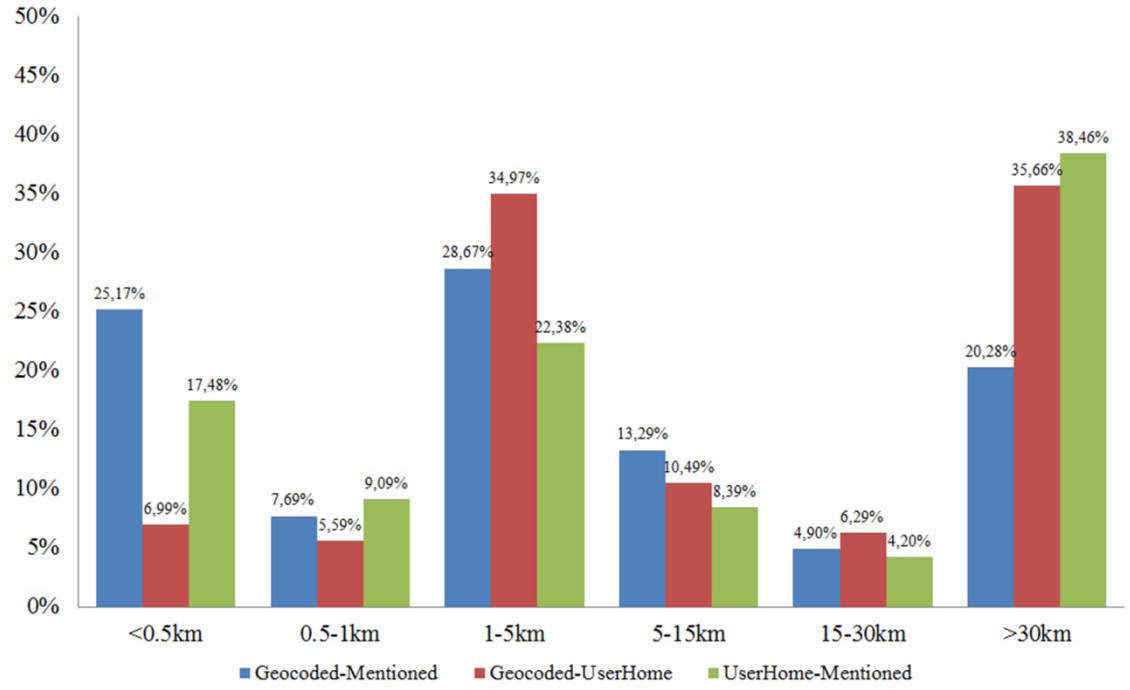


Figure 5.12: Comparative histogram with the distribution of the three spatial distances per distance range (Greater Dublin)

The histogram for the Greater Dublin dataset (Figure 5.12) shows that more than 40% of the distances related to the user home location are greater than 15 km and this finding can reinforce the conclusion that the user home location is not reliable for the urban area context applied to Dublin. Another relevant finding from the histogram shown in Figure 5.12 is that 61.53% of the tweet dataset have distances between geocoded and mentioned locations of up to 5 km, which is 1/3 of the approximated radius from a point in the city centre. This finding reinforces the reliability of the geocoded location being used instead of geoparsing the tweet message from Greater Dublin.

Similarly, the histogram for the Greater London dataset (Figure 5.13) shows that around 30% of the distances related to the user home location are greater than 30 km. Another 5% of such distances are near to the urban area limits (greater than 25 km). Thus,

it seems the user home location is not reliable for the urban area context applied to London as well.

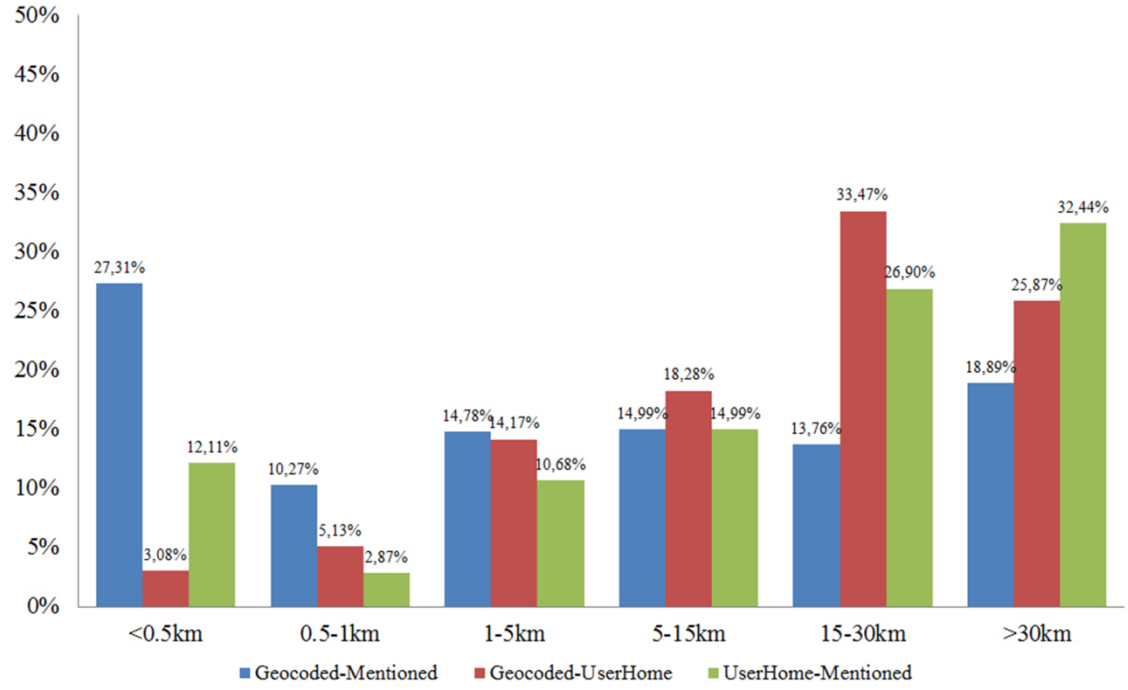


Figure 5.13: Comparative histogram with the distribution of the three spatial distances per distance range (Greater London)

Another relevant finding from the histogram shown in Figure 5.13 is that 52.36% of the tweet dataset have distances between geocoded and mentioned locations of up to 5 km, which is 1/6 of the approximate radius from a point in the city centre. In addition, 65.28 % of the dataset have such distances up to 10 km, which is 1/3 of such approximate radius. This finding also reinforces the reliability of the geocoded location on being used instead of geoparsing the tweet message from Greater London.

5.3.3 Discussion

The extended GeoSEn geoparser has presented good results on geoparsing locations mentioned in tweets from the urban GLoD. Such results suggest the gazetteer enrichment method developed on extending the geoparser is promising. However, the analysis of the relationship among tweet locations has shown that the geocoded location attached to a tweet can be used to model the spatial context of a tweet instead of performing geoparsing.

It could be noticed that the precision rates obtained by the extended GeoSEn geoparser on both Greater Dublin (68.9 %) and Greater London (66.5 %) were close to the amounts of tweets with the distance between geocoded and mentioned locations within 1/3 of the approximate radius from the cities: 61.5 % and 65.2 %, respectively. This suggests that geoparsing would only be needed on tweets without geocoding on identifying urban issues from social media or in cases where a geographical imprecision at around 1/3 of such a radius from the cities are not acceptable.

5.4 Summary

This chapter presented in details the evaluation of the research performed for this thesis. The evaluation has focused on the two main facets of the developed approach to the automated identification of urban issues from social media: the thematic and the geographical.

The evaluation of the thematic facet focused on the performance comparison of the UIDO-driven classifier with classifiers based on most commonly used machine learning algorithms adapted to process crowdsourced texts. For such, three case study setups were defined and carried out varying the datasets and the classes for classification.

The evaluation of the geographical facet focused on the performance of the extended geoparser. For such, the collected tweet datasets were geoparsed in order to find out locations mentioned in the body of tweets. Such evaluation also reported a data analysis regarding the relationship among the three tweet location contexts: the mentioned, the geocoded and the user home locations.

Finally, this chapter also presented the acquired labeled datasets from Twitter and FixMyStreet. The following chapter provides the final considerations, further work and the summary of publications generated from this thesis.

Chapter 6

Conclusion

This chapter draws some conclusions based on the results obtained along the development of this thesis. These conclusions address the research goals and answer the research questions. In addition, this chapter presents a discussion concerning proposals for further research.

6.1 Overall Discussion

This research investigated the automated identification of urban issues from social media data under two main facets: thematic and geographical. The thematic facet involved the identification and classification of urban issues. On the other hand, the geographical facet involved the identification of geographical locations in urban areas eventually mentioned in an urban issue report. The temporal facet was minimally addressed due to the complexity involving the temporal dimension and time constraints to conclude this research.

This thesis presented an innovative approach to the automated identification of Ambient Geographical Information (AGI) from social media, most specifically from the Twitter microblog, in the domain of urban issues. The proposed approach relies on ontology-driven Information Extraction to perform the identification and classification of urban issues according to semantic relationships modeled by a novel domain ontology. The

Urban Issues Domain Ontology (UIDO) was developed for structuring the vocabulary and taxonomy of urban issues. In addition, the UIDO ontology supports relationships between the thematic and geographical facets.

The geographical facet was addressed by the development of an extension of a geoparser system in order to support toponym resolution in Geographical Level of Details (GLoD) suitable for urban areas, such as Districts, Streets and Points of Interest. Such extended geoparser could then be applied to identify geographical locations eventually mentioned in non-geocoded tweets.

Concerning the thematic facet, the results obtained in comparative studies involving the UIDO ontology-driven classifier and four machine learning classifiers shown that the performance of the proposed classifier is competitive and overall satisfactory. Regarding the research question Q1, the statistical results presented in Chapter 5 shown that the null hypothesis $H1_0$ can be rejected, as the proposed classifier is competitive with all the four machine learning classifiers, especially in Setup 3, which comprises the knowledge learning from the FixMyStreet LBSN and the scenario approaches to the real world.

Regarding the research question Q2, the statistical results presented in Chapter 5 shown that the evidences were not enough to reject the null hypothesis $H2_0$. The proposed classifier performance was statistically equivalent to the two best machine learning classifiers in Setup 2, which considered the knowledge learning from manually labeled tweets in the domain of urban issues. The results suggest the need for further developments in the proposed classifier in order to reduce the number of false positives and consequently improve its performance. Moreover, further case studies with most representative datasets need to be carried out in order to make the statistical analysis stronger.

Concerning the geographical facet, the results obtained from the extended geoparser system are promising for geoparsing urban area locations mentioned in tweets. Regarding the research question Q3, the statistical results presented in Chapter 5 suggest that the null hypothesis $H3_0$ can be rejected. The analysis of the relationship among tweet location contexts shown that the geocoded location attached to a tweet is reliable to replace geographical locations eventually mentioned in a tweet regarding the context of urban areas. The precision rates obtained from the extended geoparser system in two different

urban areas were close to the percentages of tweets with the distance between geocoded and mentioned locations within 1/3 of the approximate radius from those urban areas.

Regarding the research question Q4, the statistical results presented in the analysis of manually labeled tweets in Appendix A shown that the null hypothesis $H4_0$ can be rejected. The results indicate that the sentiment polarity of a tweet is not determinative for the identification of an urban issue. This implies that the proposed approach should not rely on sentiment analysis for the identification of urban issues from social media.

Finally, this research concludes that the identification of urban issues from social media is achievable, particularly because this is the first research that addressed this problem. The research goals were fully achieved during the development of this thesis, as expected. Moreover, the proposed approach appears to be a competitive alternative to machine learning classifiers in addressing not solely the urban issues domain but also other domains that can be modeled by means of domain ontologies using knowledge from corpora or specialist knowledge.

The process of labeling tweets concerning urban issues helped to confirm that people usually complain about urban issues. Although the analyzed tweets are limited to Greater Dublin and Greater London urban areas, such finding encourages further studies in other urban areas. Such finding also reinforces the relevance of the work performed during this research. The identification of urban issues in popular streams will lead government authorities to continuously be aware of issues which are not given the required attention in LBSNs developed for such purposes. Citizens' life worldwide can be consequently improved.

6.2 Proposal for Further Research

There are some relevant topics to be addressed in order to continue the research developed in this thesis. This section enumerates and discusses some suggestions; however, interesting further research directions are not limited to these.

1. **Integrating with LBSNs in the smart cities domain**: to integrate the AGI produced by the proposed approach with LBSNs in the smart cities domain such as the Crowd4City (Falcão, 2013) and the FixMyStreet (Walravens,

2013). Such integration would enrich such platforms with urban issues reported in Twitter that have not been reported directly on them.

2. **Identifying the status of an urban issue found**: the status is the way in which the urban issue is being reported. For example, the urban issue status can be complaining (*complaint about an unsolved issue*), thanking (*reporting an action was taken to solve the issue*) and unknown (*could not recognize the user intention in the report*). Although an initial investigation found that sentiment analysis is not a determinant in identifying an urban issue, it could be helpful in this case. For example, tweets with positive sentiment polarity may indicate urban issues reported with status "thanking", indicating acknowledgments. On the other hand, tweets with negative sentiment polarity may indicate urban issues with status "complaining".

3. **Considering tweet relevance for AGI production**: to investigate ways of measuring the relevance of the tweet that produced an AGI record. For example, the tweet relevance may be higher for specific known Twitter users such as official government agencies accounts, radios and other media accounts. The tweet relevance may also be inferred from the amount of retweets, for example.

4. **Improving the UIDO ontology relationships**: exclusion rules may be included in order to improve the performance of the thematic parser. Such exclusion rules can be based on terms that are definitely not expected when identifying specific urban issue types.

5. **Enabling term disambiguation in the thematic parser**: to investigate ways of performing term-sense disambiguation in thematic candidates found. Such disambiguation would improve the accuracy of the thematic analysis as the false positive rate would decrease and urban issues would be better identified.

6. **Improving the temporal analysis in the proposed approach**: there are some known issues regarding the temporal facet that could be taken into

account. For example, how to better deal with tweets that mention two or more different timestamps or temporal periods.

7. **Supporting other text languages**: to investigate other languages such as Portuguese, Spanish, French, German, Italian, etc. Such improvement would enable the proposed approach to be applied to many different urban areas that do not have English as the official language.

8. **Using other social networks**: to integrate the proposed solution to other social networks besides Twitter, such as Facebook, Instagram, etc. For this, it is required to investigate specific APIs developed for each social network and their limitations.

9. **Improving the semantic context of a tweet**: to investigate methods and techniques that could be applied to improve the semantic context of a tweet mainly when the message is vague or contains implicit context that requires more information for inferring. For this, some ideas may be investigated, such as reading URLs mentioned in a tweet, processing previous tweets from the same user or specific hashtags. The task of reading URLs mentioned in a tweet may be helpful for both urban issue classification and toponym resolution of mentioned locations.

10. **Using knowledge from other public databases**: to investigate other LBSNs in the domain of urban issues and open databases in the Web of Data in order to extend the knowledge about urban issues and issue types. Moreover, semantic databases such as DBPedia would be helpful for improving the proposed approach.

11. **Investigating word classes with NER/NLP**: to investigate NER/NLP techniques such as post tagging in order to improve the preprocessing task and consequently the semantic analysis regarding the thematic, geographical and temporal facets.

12. **Generalizing the GeoTree structure**: to investigate ways to generalize the GeoTree structure used by the extended GeoSEn geoparser in order to classify worldwide toponyms.

13. **Automating the GLoD weights in the GeoSEn system**: to investigate ways of automating the GLoD weights used by the GeoSEn system (Campelo et al., 2009) in the confidence factors for toponym resolution. One idea for such automation is to rely on a historical analysis of occurrences in each GLoD aiming at refining the weights according to the true positive and false positive rates.

14. **Improving the GeoSEn system with tweet metadata**: to extend the GeoSEn system to use the user home location for disambiguation of geographical locations. Even though the user home location may be inaccurate and noisy, such information may be useful for the geoparser while performing disambiguation tasks. This is mainly relevant to resolve toponyms inside urban areas, which tends to present more ambiguity issues than at more general levels, such as city or state. For example, there are many streets with the same name spread around the world. For such cases, any additional knowledge about the area or even the city a tweet is related to would be decisive for the correct geographical referencing.

15. **Enriching the gazetteer with other VGI datasources**: to integrate the gazetteer enrichment strategy with other VGI datasources besides OSM, such as Foursquare, Google Places, etc. Such enrichment can improve the performance of the geoparser in urban areas.

16. **Full evaluation of the proposed approach**: to perform an evaluation comprising all the facets of the proposed approach together. Such evaluation enables to measure the performance of the entire Social2AGI system.

References

ABEL, F.; HAUFF, C.; HOUBEN, G.-J.; STRONKMAN, R.; TAO, K. Twitcident: Fighting Fire with Information from Social Web Streams. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 21., 2012, Lyon, France. **Proceedings…** New York, USA: ACM, 2012, p. 305-308.

AHLTORP, M.; TANUSHI, H.; KITAJIMA, S.; SKEPPSTEDT, M.; RZEPKA, R.; ARAKI, K. HokuMed in NTCIR-11 MedNLP-2: Automatic Extraction of Medical Complaints from Japanese Health Records Using Machine Learning and Rule-based Methods. In: NTCIR CONFERENCE ON EVALUATION OF INFORMATION ACCESS TECHNOLOGIES, 11., 2014, Tokyo, Japan. **Proceedings…** Tokyo, Japan: National Institute of Informatics, 2014, p. 158-162.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175-185, 1992.

ALVES, A. L. F.; BAPTISTA, C. S.; FIRMINO, A. A.; OLIVEIRA, M. G.; PAIVA, A. C. A Spatial and Temporal Sentiment Analysis approach applied to Twitter microtexts. **Journal of Information and Data Management**, v. 6, n. 2, p. 118-129, 2015.

AMINI, B.; IBRAHIM, R.; OTHMAN, M.S.; NEMATBAKHSH, M.A. A reference ontology for profiling scholar's background knowledge in recommender systems. Expert Systems with Applications, v. 42, p. 913-928, 2015.

ANANTHARAM, P.; BARNAGHI, P.; PRASAD, T. K.; SHETH, A. P. Extracting City Traffic Events from Social Streams. **ACM Transactions on Intelligent Systems and Technology**, v. 6, v. 4, p. 43:1-43:27, 2015.

ATEFEH, F.; KHREICH, W. A Survey of Techniques for Event Detection in Twitter. **Computational Intelligence**, v. 31, n. 1, p. 132-164, 2013.

AUGUSTINE, E.; CUSHING, C.; DEKHTYAR, A.; MCENTEE, K.; PATERSON, K.; TOGNETTI, M. Outage Detection via Real-time Social Stream Analysis: Leveraging the Power of Online Complaints. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 21., 2012, Lyon, France. **Proceedings…** New York, USA: ACM, 2012, p. 13-22.

BALLATORE, A.; BERTOLOTTO, M. Semantically enriching VGI in support of implicit feedback analysis. In: TANAKA, K.; FRÖHLICH, P.; KIM, K-S. (Ed.). **Web and Wireless Geographical Information Systems (LNCS)**, v. 6574. Kyoto, Japan: Springer, 2011, p. 78-93.

BALLATORE, A.; WILSON, D. C; BERTOLOTTO, M. Computing the semantic similarity of geographic terms using volunteered lexical definitions. **International Journal of Geographical Information Science**, v. 27, n. 10, p. 2099-2118, 2013.

BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. **Modern Information Retrieval**. ACM Press/Addison-Wesley, 1999.

BEARD, K. A Semantic Web Based Gazetteer Model for VGI. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON CROWDSOURCED AND VOLUNTEERED GEOGRAPHIC INFORMATION, 1., 2012, Redondo Beach, USA. **Proceedings…** New York, USA: ACM, 2012, p. 54-61.

BOMMANNAVAR, P.; LIN, J.; RAJARAMAN. A. Estimating topical volume in social media streams. In: Annual ACM Symposium on Applied Computing, 31., 2016, Pisa, Italy. **Proceedings…** New York, USA: ACM, 2016, p. 1096-1101.

BORDOGNA, G.; GHISALBERTI, G.; PSAILA, G. Geographic information retrieval: Modeling uncertainty of user's context. **Fuzzy Sets and Systems**, v. 196, p. 105-124, 2012.

BOULIS, C.; OSTENDOF, M. Text Classification by Augmenting the Bag-of-Words Representation with Redundancy-Compensated Bigrams. In: SIAM International Conference on Data Mining at the Workshop on Feature Selection in Data Mining, 1., 2005, Newport Beach, USA. **Proceedings…** Philadelphia, USA: SIAM, 2005.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

CAMPELO, C. E. C. **GeoSEn: a search engine with geographical focus** (in Portuguese). Campina Grande, Brazil: UFCG, 2008. Master Thesis, Federal University of Campina Grande, Brazil, 2008, 150 p.

CAMPELO, C. E. C.; BAPTISTA, C. S. A Model for Geographic Knowledge Extraction on Web Documents. In: HEUSER, C. A.; PERNUL, G. (Ed.). **Advances in Conceptual Modeling - Challenging Perspectives (LNCS)**, v. 5833. Gramado, Brazil: Springer, 2009, p. 317-326.

CARAGLIU, A.; DEL BO, C.; NIJKAMP, P. Smart cities in Europe. **Journal of Urban Technology**, v. 18, n. 2, p. 65-82, 2011.

CARDOSO, S. D.; AMANQUI, F. K.; SERIQUE, K. J. A.; DOS SANTOS, J. L. C.; MOREIRA, D. A. SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data. **Future Generation Computer Systems**, v.54, n. 1, p. 389-398, 2016.

CHAN, C.K.; VASARDANI, M.; WINTER, S. Leveraging Twitter to detect event names associated with a place. **Journal of Spatial Science**, v. 59, n. 1, p. 137-155, 2014.

CHANG, A.X; MANNING, C.D. SUTIME: A Library for Recognizing and Normalizing Time Expressions. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8., 2012, Istanbul, Turkey. **Proceedings…** Paris, France: European Language Resources Association, 2012, p. 3735-3740.

CHANG, C-C; LIN, C-J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, n. 3, p. 27:1-27:27, 2011.

CHENLIANG, L.; WENG, J.; HE, Q.; YAO, Y.; DATTA, A.; SUN, A.; LEE, B-S. TwiNER: named entity recognition in targeted twitter stream. In: INTERNATIONAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 35., 2012, Portland, USA. **Proceedings…** New York, USA: ACM, 2012, p. 721-730.

CIMIANO, P.; VOLKER, J. Text2Onto - A framework for ontology learning and data-driven change discovery. In: INTERNATIONAL CONFERENCE ON NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, 10., 2005, Alicante, Spain. **Proceedings…** Heidelberg, Germany: Springer, 2005, p. 227-238.

CORCHO, O. Ontology based Document Annotation: Trends and Open Research Problems. **Journal of Metadata, Semantics and Ontologies**, v. 1, n. 1, p. 47-57, 2006.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273-297, 1995.

CRANSHAW, J.; SCHWARTZ, R.; HONG, J. I.; SADEH, N. M. The Livehoods Project: utilizing social media to understand the dynamics of a city. In: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 6., 2012, Dublin, Ireland. **Proceedings…** Palo Alto, USA: The AAAI Press, 2012, p. 58-65.

CRANSHAW, J.; TOCH, E.; HONG, J.; KITTUR, A.; SADEH, N. Bridging the gap between physical location and online social networks. In: ACM INTERNATIONAL CONFERENCE ON UBIQUITOUS COMPUTING, 12., 2010, Copenhagen, Denmark. **Proceedings…** New York, USA: ACM, 2010, p. 119-128.

CROOKS, A.; PFOSER, D.; JENKINS, A.; CROITORU, A.; STEFANIDIS, A.; SMITH, D.; KARAGIORGOU, S.; EFENTAKIS, A.; LAMPRIANIDIS, G. Crowdsourcing urban form and function. **International Journal of Geographical Information Science**, v. 29, n. 5, p. 720-741, 2015.

CUNNINGHAM, H.; TABLAN, V.; ROBERTS, A.; BONTCHEVA, K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. **PLOS Computational Biology**, v. 9, n. 2, p. 1-16, 2013.

DEL BIMBO, A.; FERRACANI, A.; PEZZATINI, D.; D'AMATO, F.; SERENI, M. LiveCities: Revealing the Pulse of Cities by Location-based Social Networks Venues and Users Analysis. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 23., 2014, Seoul, Korea. **Proceedings…** Geneva, Switzerland: International WWW Conferences Steering Committee, 2014, p. 163-166.

DÉSIR, C.; BERNARD, S.; PETITJEAN, C.; LAURENT, H. One class random forests. **Pattern Recognition**, v. 46, n. 12, p. 3490-3506, 2013.

EL-BELTAGY, S. R.; RAFEA, A. KP-Miner: A keyphrase extraction system for English and Arabic documents. **Information Systems**, v. 34, n. 1, p. 132-144, 2009.

EMBLEY, D. W.; CAMPBELL, D. M., SMITH, R.D. Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 7., 1998, Washington, USA. **Proceedings…** New York, USA: ACM, 1998, p. 52-59.

ESHLEMAN, R.; YANG, H. Hey #311, Come Clean My Street!: A Spatio-temporal Sentiment Analysis of Twitter Data and 311 Civil Complaints. In: IEEE INTERNATIONAL CONFERENCE ON BIG DATA AND CLOUD COMPUTING, 4., 2014, Sydney, Australia. **Proceedings…** Washington, USA: IEEE Computer Society, 2014, p. 477-484.

FALCÃO, A. G. R. **Crowd4City: a location-based social network applied to the smart-cities domain** (in Portuguese). Campina Grande, Brazil: UFCG, 2013. Master Thesis, Federal University of Campina Grande, Brazil, 2013, 84 p.

FERNÁNDEZ-LÓPEZ, M.; GÓMEZ-PÉREZ, A. Overview and analysis of methodologies for building ontologies. **The Knowledge Engineering Review**, v. 17, n. 2, p. 129-156, 2002.

FEYISETAN, O.; SIMPERL, E.; TINATI, R.; LUCZAK-ROESCH, M.; SHADBOLT, N. Quick-and-clean Extraction of Linked Data Entities from Microblogs. In: INTERNATIONAL CONFERENCE ON SEMANTIC SYSTEMS, 10., 2014, Leipzig, Germany. **Proceedings…** New York, USA: ACM, 2014, p. 5-12.

FONSECA, F.; DAVIS, C.; CAMARA, G. Bridging ontologies and conceptual schemas in geographic applications development. **Geoinformatica**, v. 7, n. 4, p. 355-378, 2003.

FREITAG, D. Machine Learning for Information Extraction in Informal Domains. **Machine Learning**: special issue on Information Retrieval, v. 39, n. 2-3, p. 169-202, 2000.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian Network Classifiers. **Machine Learning**. v. 29, n. 2-3, p. 131-163, 1997.

FURTADO, V.; CAMINHA, C.; AYRES, L.; SANTOS, H. Open Government and Citizen Participation in Law Enforcement via Crowd Mapping. **IEEE Intelligent Systems**, v. 27, n. 4, p. 63-69, 2012.

GARCIA, A.; CAMACHO, C.; BELLENZIER, M.; PASQUALI, M.; WEBER, T.; SILVEIRA, M. S. Data Visualization in Mobile Applications: Investigating a Smart City App. In: KUROSU, M. (Ed.). **Human-Computer Interaction. Interaction Platforms and Techniques (LNCS)**, v. 9732. Toronto, Canada: Springer, 2016, p. 285-293.

GELERNTER, J.; GANESH, G.; KRISHNAKUMAR, H.; ZHANG, W. Automatic Gazetteer Enrichment with User-geocoded Data. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON CROWDSOURCED AND VOLUNTEERED GEOGRAPHIC INFORMATION, 2., 2013, Orlando, USA. **Proceedings…** New York, USA: ACM, 2013, p. 87-94.

GELERNTER, J.; MUSHEGIAN, N. Geo-parsing Messages from Microtext. **Transactions in GIS**, v. 15, n. 6, p. 753-773, 2011.

GOODCHILD, M. F. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. **International Journal of Spatial Data Infrastructures Research**, v. 2, p. 24-32, 2007.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**. v. 5, n. 2, p. 199-220, 1993.

GRUNINGER, M.; LEE, J. Ontology applications and design. **Communications of the ACM**, v. 45, n. 2, p. 39-41, 2002.

HAGHIGHI, P. D.; BURSTEIN, F.; ZASLAVSKY, A.; ARBON, P. Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings. **Decision Support Systems**, v. 54, n. 2, p. 1192-1204, 2013.

HAKLAY, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. **Environment and Planning B: Planning and Design**, v. 37, n. 4, p. 682-703, 2010.

HAKLAY, M.; WEBER, P. OpenStreetMap: User-Generated Street Maps. **IEEE Pervasive Computing**, v. 7, n. 4, p. 12-18, 2008.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B. REUTEMANN, P.; WITTEN, I.H. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v. 11, n. 1, p. 10-18, 2009.

HAVLIK, D.; SORIANO, J.; GRANELL, C.; MIDDLETON, S. E.; VAN DER SCHAAF, H.; BERRE, A. J.; PIELORZ, J. Future Internet enablers for VGI applications. In: ENVIRONMENTAL INFORMATICS AND RENEWABLE ENERGIES, 27., 2013, Hamburg, Germany. **Proceedings…** Hamburg: Universität Hamburg, 2013, p. 622-630.

HAWELKA, B.; SITKO, I.; BEINAT, E.; SOBOLEVSKY, S.; KAZAKOPOULOS, P.; RATTI, C. Geo-located twitter as proxy for global mobility patterns. **Cartography and Geographic Information Science**, v. 41, n. 3, p. 260-271, 2014.

HE, J.; HU, M.; SHI, M.; LIU, Y. Research on the measure method of complaint theme influence on online social network. **Expert Systems with Applications**, v. 41, n. 13, p. 6039-6046, 2014.

HORITA, F. E. A.; DEGROSSI, L. C.; ASSIS, L. F. G.; ZIPF, A.; ALBUQUERQUE, J. P. The use of Volunteered Geographic Information (VGI) and Crowdsourcing in Disaster Management: a Systematic Literature Review. In: AMERICAS CONFERENCE ON INFORMATION SYSTEMS, 19., 2013, Chicago, USA. **Proceedings…** Chicago: AIS, 2013, p. 1-10.

IMRAN, M.; CASTILLO, C.; LUCAS, J.; PATRICK, M.; ROGSTADIUS, J. Coordinating human and machine intelligence to classify microblog communications in crises. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS FOR CRISIS RESPONSE AND MANAGEMENT, 11., 2014, State College, USA. **Proceedings…** State College, USA: The Pennsylvania State University, 2014, p. 712-721.

JACKSON, M. J.; RAHEMTULLA, H.; MORLEY, J. The Synergistic use of authenticated and crowd-sourced data for Emergency response. In: INTERNATIONAL WORKSHOP ON VALIDATION OF GEO-INFORMATION PRODUCTS FOR CRISIS MANAGEMENT, 2., 2010, Ispra, Italy. **Proceedings…** Brussels, Belgium : JRC, 2010, p. 91-99.

JARRAR, M. Towards Effectiveness and Transparency in e-Business Transactions, An Ontology for Customer Complaint Management. In: GARCIA, R. (Ed.). **Semantic Web for Business: Cases and Applications**. Hershey, USA: IGI Global, 2009. chap 7, p. 127-149.

JIN, J.; YAN, X.; YU, Y.; LI, Y. Service Failure Complaints Identification in Social Media: A Text Classification Approach. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 34., 2013, Milan, Italy. **Proceedings…** Milan: AIS, 2013, p. 1-11.

JOHN, G.H.; LANGLEY, P. Estimating Continuous Distributions in Bayesian Classifiers. In: Conference on Uncertainty in Artificial Intelligence, 11., San Mateo, USA. **Proceedings…** San Francisco, USA: Morgan Kaufmann Publishers Inc., 1995, p. 338-345.

JUNG, J. J. Towards Named Entity Recognition Method for Microtexts in Online Social Networks: A Case Study of Twitter. In: INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING, 2., 2011, Kaohsiung, Taiwan. **Proceedings…** Washington, USA: IEEE Computer Society, 2011, p. 563-564.

KANG, Y.-B.; HAGHIGHI, P. D.; BURSTEIN, F. CFinder: An Intelligent Key Concept Finder from Text for Ontology Development. **Expert Systems with Applications**, v. 41, n. 9, p. 4494-4504, 2014.

KEβLER, C.; JANOWICZ, K.; BISHR, M. An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In: ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 17., 2009, Seattle, USA. **Proceedings…** New York, USA: ACM, 2009, p. 91-100.

KITCHIN, R. The real-time city? Big data and smart urbanism. **GeoJournal**, v. 79, n. 1, p. 1-14, 2014.

KITCHIN, R.; LAURIAULT, T. P.; MCARDLE, G. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. **Regional Studies, Regional Science**, v. 2, n. 1, p. 6-28, 2015.

KISH, L. **Statistical Design for Research**. New Jersey, USA: Wiley, 2004.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence, 14., 1995, Montreal, Canada. **Proceedings…** San Francisco, USA: Morgan Kaufmann Publishers Inc., 1995, p. 1137-1143.

KOULOUMPIS, E.; WILSON, T.; MOORE, J. Twitter Sentiment Analysis: The Good the Bad and the OMG!. In: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 5., 2011, Barcelona, Spain. **Proceedings…** Menlo Park, USA: The AAAI Press, 2011, p. 538-541.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 18., 2001, Williamstown, USA. **Proceedings…** San Francisco, USA: Morgan Kaufmann, 2001, p. 282-289.

LAMPRIANIDIS, G.; SKOUTAS, D.; PAPATHEODOROU, G.; PFOSER, D. Extraction, Integration and Analysis of Crowdsourced Points of Interest from Multiple Web Sources. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON CROWDSOURCED AND VOLUNTEERED GEOGRAPHIC INFORMATION, 3., 2014, Dallas, USA. **Proceedings…** New York, USA: ACM, 2014, p. 16-23.

LEE, C.-H.; WANG, Y.-H.; TRAPPEY, A. J. C. Ontology-based reasoning for the intelligent handling of customer complaints. **Computers & Industrial Engineering**, v. 84, n. 1, p. 144-155, 2015.

LEE, R.; WAKAMIYA, S.; SUMIYA, K. Urban area characterization based on crowd behavioral lifelogs over Twitter. **Personal and Ubiquitous Computing**, v. 17, n. 4, p. 605-620, 2013.

LEETARU, K.; WANG, S.; CAO, G.; PADMANABHAN, A.; SHOOK, E. Mapping the global Twitter heartbeat: The geography of Twitter. **First Monday, [S.l.]**, 2013. ISSN 13960466. Available at: <http://journals.uic.edu/ojs/index.php/fm/article/view/4366/3654>. Last access: 17 nov 2016.

LI, C.; SUN, A. Fine-grained Location Extraction from Tweets with Temporal Awareness. In: INTERNATIONAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 37., 2014, Queensland, Australia. **Proceedings…** New York, USA: ACM, 2014, p. 43-52.

LIANG, Y.; CAVERLEE, J.; CHENG, Z.; KAMATH, K. Y. How Big is the Crowd? Event and Location Based Population Modeling in Social Media. In: ACM CONFERENCE ON HYPERTEXT AND SOCIAL MEDIA, 24., 2013, Paris, France. **Proceedings…** New York, USA: ACM, 2013, p. 99-108.

LINGAD, J.; KARIMI, S.; YIN, J. Location Extraction from Disaster-related Microblogs. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 22., 2013, Rio de Janeiro, Brazil. **Proceedings…** Geneva, Switzerland: International WWW Conferences Steering Committee, 2013, p. 1017-1020.

LIU, Y.; KLIMAN-SILVER, C.; MISLOVE, A. The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. In: INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 8., 2014, Ann Arbor, USA. **Proceedings…** Palo Alto, USA: The AAAI Press, 2014, p. 305-314.

MAGDY, A.; ALARABI, L.; AL-HARTHI, S.; MUSLEH, M.; GHANEM, T. M.; GHANI, S.; MOKBEL, M. F. Taghreed: a system for querying, analyzing, and visualizing geotagged microblogs. In: ACM SIGSPATIAL INTERNATIONAL CONFERENCE ON ADVANCES IN GEOGRAPHIC INFORMATION SYSTEMS, 22., Dallas, USA. **Proceedings…** New York, USA: ACM, 2014, p. 163-172.

MAHMUD, J.; NICHOLS, J.; DREWS, C. Home Location Identification of Twitter Users. **ACM Transactions on Intelligent Systems and Technology**: special section on Urban Computing, v. 5, n. 3, p. 47:1-47:21, 2014.

MARTINS, B.; SILVA, M. J.; CHAVES, M. S. Challenges and Resources for Evaluating Geographical IR. In: WORKSHOP ON GEOGRAPHIC INFORMATION RETRIEVAL, 2., 2005, Bremen, Germany. **Proceedings…** New York, USA: ACM, 2005, p. 65-69.

MCCALLUM, A.; NIGAM, K. A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI Workshop on Learning for Text Categorization, 1998, Madison, USA. **Proceedings…** Dortmund, Germany: Universitaet Dortmund, 1998, p. 41-48.

MCGUINNESS, D. L.; VAN HARMELEN, F. **OWL Web Ontology Language overview**. W3C recommendation 10 February 2004, World Wide Web Consortium (W3C), 2004. Available at: <http://www.w3.org/TR/owl-ref/>. Last access: 10 dec 2015.

METZLER, D. **A Feature-Centric View of Information Retrieval**. The Information Retrieval Series, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

MILLER, G. A. WordNet: A Lexical Database for English. **Communications of the ACM**, v. 38, n. 11, p. 39-41, 1995.

MOONEY, P.; CORCORAN, P. The Annotation Process in OpenStreetMap. **Transactions in GIS**, v. 16, n. 4, p. 561-579, 2012.

MOURA, T. H. V. M., DAVIS JR., C. A. Integration of linked data sources for gazetteer expansion. In: ACM SIGSPATIAL WORKSHOP ON GEOGRAPHIC INFORMATION RETRIEVAL, 8., 2014, Dallas, USA. **Proceedings…** New York, USA: ACM, 2014, p. 5:1-5:8.

NAKHASI, A.; PASSARELLA, R.; BELL, S.G.; PAUL, M.J.; DREDZE, M.; PRONOVOST, P. Malpractice and Malcontent: Analyzing Medical Complaints in Twitter. In: AAAI FALL SYMPOSIUM SERIES: INFORMATION RETRIEVAL AND KNOWLEDGE DISCOVERY IN BIOMEDICAL TEXT, 21., 2012, Arlington, USA. **Proceedings…** Palo Alto, USA: The AAAI Press, 2012, p. 84-85.

NGO, A.; REVESZ, P. Efficient Traffic Crash and Snow Complaint GIS System. In: ANNUAL INTERNATIONAL DIGITAL GOVERNMENT RESEARCH CONFERENCE: DIGITAL GOVERNMENT INNOVATION IN CHALLENGING TIMES, 12., 2011, College Park, USA. **Proceedings…** New York, USA: ACM, 2011, p. 235-244.

NIST/SEMATECH. Exponential Smoothing. In: _____. **e-Handbook of Statistical Methods**, section 6.4.3. Gaithersburg, USA, 2012. Available at: <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc43.htm>. Last access: 6 jan 2016.

OGC; W3C. **Time Ontology in OWL**. 2016. Available at: <https://www.w3.org/TR/owl-time/>. Last access: 12 sep 2016.

OLIVEIRA, M. G.; BAPTISTA, C. S.; CAMPELO, C. E. C.; ACIOLI FILHO, J. A. M.; FALCÃO, A. G. R. Automated Production of Volunteered Geographic Information from Social Media. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 15., 2014, Campos do Jordão, Brazil. **Proceedings…** São José dos Campos: MCTI/INPE, 2014, p. 118-129.

OLIVEIRA, M. G.; BAPTISTA, C. S.; CAMPELO, C. E. C.; ACIOLI FILHO, J. A. M.; FALCÃO, A. G. R. Producing Volunteered Geographic Information from Social Media for LBSN Improvement. **Journal of Information and Data Management**, v. 6, n. 1, p. 81-91, 2015 (a).

OLIVEIRA, M. G.; CAMPELO, C. E. C.; BAPTISTA, C. S.; BERTOLOTTO, M. Leveraging VGI for Gazetteer Enrichment: a case study for geoparsing twitter messages. In: GENSEL, J.; TOMKO, M. (Ed.). **Web and Wireless Geographical Information Systems (LNCS)**, v. 9080. Grenoble, France: Springer, 2015 (b), p. 20-36.

OLIVEIRA, M. G.; CAMPELO, C. E. C.; BAPTISTA, C. S.; BERTOLOTTO, M. Gazetteer enrichment for addressing urban areas: a case study. **Journal of Location Based Services**, v. 10, n. 2, p. 142-159, 2016.

ONORATI, T.; MALIZIA, A.; DIAZ, P.; AEDO, I. Modeling an ontology on accessible evacuation routes for emergencies. **Expert Systems with Applications**, v. 41, p. 7124-7134, 2014.

PARKER, C. J. **The Fundamentals of Human Factors Design for Volunteered Geographic Information**, London, UK: Springer, 2014.

PARKER, C. J.; MAY, A. J.; MITCHELL, V. Relevance of volunteered geographic information in a real world context. In: GEOGRAPHICAL INFORMATION SCIENCE RESEARCH UK, 19., 2011, Portsmouth, UK. **Proceedings…** Portsmouth, UK: University of Portsmouth, 2011.

PETERS, I. **Folksonomies. Indexing and Retrieval in Web 2.0**, Berlin, GE: De Gruyter Saur, 2009.

PORTER, M. F. An algorithm for suffix stripping. In: JONES, K. S.; WILLETT, P. (Ed.). **Readings in Information Retrieval**, v. 1, San Francisco, USA: Morgan Kaufmann Publishers, 1997, p. 313-316.

PURVES, R.; JONES, C. Geographic Information Retrieval. **SIGSPATIAL Special**, v. 3, n. 2, p. 2-4, 2011.

QUINLAN, R. **C4.5: Programs for Machine Learning**. San Mateo, USA: Morgan Kaufmann Publishers, 1993.

RAIMOND, Y.; ABDALLAH, S. **The Event Ontology**, 2007. Available at: <http://motools.sf.net/event/event.html>. Last access: 26 jan 2016.

RATINOV, L.; ROTH, D. Design Challenges and Misconceptions in Named Entity Recognition. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 13., 2009, Boulder, USA. **Proceedings…** Stroudsburg, USA: Association for Computational Linguistics, 2009, p. 147-155.

RITTER, A.; CLARK, S.; MAUSAM; ETZIONI, O. Named Entity Recognition in Tweets: An Experimental Study. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 16., 2011, Edinburgh, Scotland. **Proceedings…** Stroudsburg, USA: Association for Computational Linguistics, 2011, p. 1524-1534.

ROUSSEY, C.; PINET, F.; KANG, M.A.; CORCHO, O. An Introduction to Ontologies and Ontology Engineering. In: FALQUET, G. et al. (Ed). **Ontologies in Urban Development Projects (AIKP)**, v. 1. Springer, 2011, p. 9-38.

SAKAKI, T.; OKAZAKI, M.; MATSUO, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 19., 2010, Hong Kong. **Proceedings…** New York, USA: ACM, 2010, p. 851-860.

SHAPIRO, S.S.; WILK, M.B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3-4, p. 591-611, 1965.

SIDAWY, E. **An application to inform cities of public space damaging**. INNOV'in the city, 2010. Available at: <http://www.innovcity.com/2010/05/29/an-application-to-inform-cities-of-public-space-damaging/>. Last access: 19 jan 2016.

SIDOROV, G.; GELBUKH, A.; GÓMEZ-ADORNO, H.; PINTO, D. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. **Computación y Sistemas**, v. 18, n. 3, p. 491-504, 2014.

SIMPERL, E. Reusing ontologies on the Semantic Web: A feasibility study. **Data & Knowledge Engineering**, v. 68, n. 10, p. 905-925, 2009.

SPÄRCK JONES, K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. **Journal of Documentation**, v. 28, p. 11-21, 1972.

SPINSANTI, L.; OSTERMANN, F. Automated geographic context analysis for volunteered information. **Applied Geography**, v. 43, n. 1, p. 36-44, 2013.

STADLER, C.; LEHMANN, J.; HÖFFNER, K.; AUER, S. LinkedGeoData: A core for a web of spatial open data. **Semantic Web Journal**, v. 3, n. 4, p. 333-354, 2012.

STEFANIDIS, A.; CROOKS, A.; RADZIKOWSKI, J. Harvesting ambient geospatial information from social media feeds. **GeoJournal**, v. 78, n. 2, p. 319-338, 2013.

SUROWIECKI, J. **The Wisdom of Crowds**. New York, USA: Anchor, 2005.

THRIFT, N. The promise of urban informatics: some speculations. **Environment and Planning A**, v. 46, n. 6, p. 1263-1266, 2014.

TONELLI, S.; ROSPOCHER, M.; PIANTA, E.; SERAFINI, L. Boosting collaborative ontology building with key-concept extraction. In: IEEE INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, 5., 2011, Palo Alto, USA. **Proceedings…** Palo Alto: IEEE Computer Society, 2011, p. 316-319.

TRIM, C. **TF/IDF with Google n-Grams and POS Tags**, 2013. Available at: <http://trimc-nlp.blogspot.ie/2013/04/tfidf-with-google-n-grams-and-pos-tags.html>. Last access: 25 jul 2016.

TSAMPOULATIDIS, I.; VERVERIDIS, D.; TSARCHOPOULOS, P.; NIKOLOPOULOS, S.; KOMPATSIARIS, I.; KOMNINOS, N. ImproveMyCity: An Open Source Platform for Direct Citizen-government Communication. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 21., 2013, Barcelona, Spain. **Proceedings…** New York, USA: ACM, 2013, p. 839-842.

WAKAMIYA, S.; BELOUAER, L.; BROSSET, D.; LEE, R.; KAWAI, Y.; SUMIYA, K.; CLARAMUNT, C. Measuring Crowd Mood in City Space Through Twitter. In: GENSEL, J.; TOMKO, M. (Ed.). **Web and Wireless Geographical Information Systems (LNCS)**, v. 9080. Grenoble, France: Springer, 2015, p. 37-49.

WAKAMIYA, S.; LEE, R.; SUMIYA, K. Crowd-sourced urban life monitoring: Urban area characterization based crowd behavioral patterns from Twitter. In: INTERNATIONAL CONFERENCE ON UBIQUITOUS INFORMATION MANAGEMENT AND COMMUNICATION, 6., 2012, Danang, Vietnam. **Proceedings…** New York, USA: ACM, 2012, Article no. 26.

WALLGRÜN, J. O.; HARDISTY, F.; MACEACHREN, A. M.; KARIMZADEH, M.; JU, Y.; PEZANOWSKI, S. Construction and First Analysis of a Corpus for the Evaluation and Training of Microblog/Twitter Geoparsers. In: ACM SIGSPATIAL WORKSHOP ON GEOGRAPHIC INFORMATION RETRIEVAL, 8., 2014, Dallas, USA. **Proceedings…** New York, USA: ACM, 2014, p. 4:1-4:8.

WALRAVENS, N. Validating a Business Model Framework for Smart City Services: The Case of FixMyStreet. In: INTERNATIONAL CONFERENCE ON ADVANCED INFORMATION NETWORKING AND APPLICATIONS WORKSHOPS, 27., 2013, Barcelona, Spain. **Proceedings…** Los Alamitos, USA: IEEE Computer Society, 2013, p. 1355-1360.

WANG, W.; STEWART, K. Spatiotemporal and semantic information extraction from Web news reports about natural hazards. **Computers, Environment and Urban Systems**, v. 50, n. 1, p. 30-40, 2015.

WANNER, L.; ROSPOCHER, M.; VROCHIDIS, S.; JOHANSSON, L.; BOUAYAD-AGHA, N.; CASAMAYOR, G.; KARPPINEN, A.; KOMPATSIARIS, I.; MILLE, S.; MOUMTZIDOU, A.; SERAFINI, L. Ontology-centered environmental information delivery for personalized decision support. **Expert Systems with Applications**, v. 42, p. 5032-5046, 2015.

WATANABE, K.; OCHI, M.; OKABE, M.; ONAI, R. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In: ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 20., Glasgow, Scotland. **Proceedings…** New York, USA: ACM, 2011, p. 2541-2544.

WILCOXON, F. Individual comparisons by ranking methods. **Biometrics Bulletin**, v. 1, n. 6, p. 80-83, 1945.

WIMALASURIYA, D. C.; DOU, D. Ontology-based information extraction: An introduction and a survey of current approaches. **Journal of Information Science**, v. 36, n. 3, p. 306-323, 2010.

WINKLER, W.E. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. In: **Section on Survey Research**, p. 354-359, 1990.

WU, H. C.; LUK, R. W. P.; WONG, K. F.; KWOK, K. L. Interpreting TF-IDF Term Weights As Making Relevance Decisions. **ACM Transactions on Information Systems**, v. 26, n. 3, p. 13:1–13:37, 2008.

XIA, C.; HU, J.; ZHU, Y.; NAAMAN, M. What Is New in Our City? A Framework for Event Extraction Using Social Media Posts. In: CAO, T.; LIM, E.-P.; ZHOU, Z.-H.; HO, T.-B.; CHEUNG, D.; MOTODA, H. (Ed.). **Advances in Knowledge Discovery and Data Mining (LNCS)**, v. 9070, Ho Chi Minh City, Vietnam: Springer, 2015, p. 16-32.

XU, Z.; ZHANG, H.; LIU, Y.; MEI, L. Crowd Sensing of Urban Emergency Events Based on Social Media Big Data. In: IEEE INTERNATIONAL CONFERENCE ON TRUST, SECURITY AND PRIVACY IN COMPUTING AND COMMUNICATIONS, 13., 2014, Beijing, China. **Proceedings…** Washington, USA: IEEE Computer Society, 2014, p. 605-610.

YANG, M.; LI, Y.; KIANG, M. Environmental Scanning for Customer Complaint Identification in Social Media. In: INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS, 32., 2011, Shanghai, China. **Proceedings…** Shanghai: AIS, 2011.

YANG, X.; GAO, R.; HAN, Z.; SUI, X. Ontology-Based Hazard Information Extraction from Chinese Food Complaint Documents. In: TAN, Y.; SHI, Y.; JI, Z. (Ed.). **Advances in Swarm Intelligence (LNCS)**, v. 7332, Shenzhen, China: Springer, 2012, p. 155-163.

YE, M.; YIN, Y.; LEE, W. Q.; LEE, D. L. Exploiting geographical influence for collaborative point-of-interest recommendation. In: INTERNATIONAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 34., 2011, Beijing, China. **Proceedings…** New York, USA: ACM, 2011, p. 325-334.

YIN, J.; KARIMI, S.; LINGAD, J. Pinpointing Locational Focus in Microblogs. In: AUSTRALASIAN DOCUMENT COMPUTING SYMPOSIUM, 19., 2014, Melbourne, Australia. **Proceedings…** New York, USA: ACM, 2014, p. 66:66-66:72.

YUAN, N. J.; ZHANG, F.; LIAN, D.; ZHENG, K.; YU, S.; XIE, X. We know how you live: Exploring the spectrum of urban lifestyles. In: ACM CONFERENCE ON ONLINE SOCIAL NETWORKS, 1., 2013, Boston, USA. **Proceedings…** New York, USA: ACM, 2013, p. 3-14.

ZHENG, Y.; CAPRA, L.; WOLFSON, O.; YANG, H. Urban Computing: Concepts, Methodologies, and Applications. **ACM Transactions on Intelligent Systems and Technology**: special section on Urban Computing, v. 5, n.3, p. 38:1–38:55, 2014 (a).

ZHENG, Y.; LIU, T.; WANG, Y.; ZHU, Y.; LIU, Y.; CHANG, E. Diagnosing New York City's Noises with Ubiquitous Data. In: ACM INTERNATIONAL JOINT CONFERENCE ON PERVASIVE AND UBIQUITOUS COMPUTING, SEATTLE, USA. **Proceedings…** New York, USA: ACM, 2014 (b), p. 715-725.

ZIRTILOĞLU, H.; YOLUM, P. Ranking Semantic Information for e-Government: Complaints Management. In: INTERNATIONAL WORKSHOP ON ONTOLOGY-SUPPORTED BUSINESS INTELLIGENCE, 1., 2008, Karlsruhe, Germany. **Proceedings…** New York, USA: ACM, 2008, pp. 5:1–5:7

Appendix A

Labeling Tweets on Urban Issues and Urban Places

This appendix depicts in details the process performed for building a dataset of labeled tweets in the domain of urban issues for the evaluation of the thematic facet in the proposed approach to automated identification of urban issues from social media. In addition, urban place names mentioned in the tweets were also identified and labeled in order to be used for the evaluation of the geographical facet.

The lack of manually annotated corpora for evaluation and training of Twitter geoparsers is discussed by Wallgrün et al. (2014). They have presented an approach in order to provide an English corpus of manually geocoded tweets, focusing solely on the spatial dimension and with GLoD not suitable for urban places, such as cities and countries. Based on such ideas for building a corpus of labeled tweets, the following subsections depicts the details of the developed approach to labeling tweets suitable for the scope of this research.

A.1 Harvesting

Aiming at using real data for the evaluation of the proposed approach on this thesis, a crawler system was developed in Java language. Such crawler ran continuously,

crawling tweets every minute. The crawler focused on four different urban areas around the world: London, UK; Dublin, Ireland; Campina Grande, Brazil; and Maceió, Brazil; and two different speaking languages: English and Portuguese.

In order to harvest tweets from these urban areas, the crawler relies on the Twitter Streamming API[67] through the geographical search provided. The idea behind collecting solely geocoded tweets is to provide a corpus from a homogeneous urban area, language, and culture. Since Twitter search APIs are very limited on performing historical searches, allowing only searches on tweets posted in the last seven days, the crawler kept running and collecting tweets in real time from the stream. However, the streaming Twitter API allows retrieving at most 1% of the Tweet feed per moment (Liu et al., 2014).

The geographical search provided by such API requires solely the bounding box from the search area in Comma Separated Values (CSV) format. Unfortunately, such API does not support polygon-based geographical queries. Thus, the Klokantech BoundingBox[68] Web application was used to collect the bounding boxes from the four urban areas selected. Each tweet was retrieved in a JSON file. These JSON files were stored as a text field in a PostgreSQL local database to being further processed.

A simply table template was developed with only three columns: `tid` (*serial*), a local unique identifier of the tweet; `crawler_datetime` (*timestamp with timezone*), the datetime the tweet was collected; and `json` (*text*), which stores the JSON file as a text string. Thus, each tweet was stored as a record in such table. There is one different table using the sample template for each different city. Such simply table structure was developed in order to collect the stream of tweets with a minimum latency, avoiding losing tweets from the streaming. One example of a JSON file containing a tweet is depicted in Appendix B. Table A.1 describes the four datasets of tweets harvested during this research.

As it can be seen in Table A.1, the amount of tweets is too large to be manually classified by humans. This fact implies to select samples from those datasets in order to perform the labeling, as stated by Kish (2004). For such, a filtering strategy was developed and applied. As this research focused solely on English tweets, the following steps were applied only in the tweet datasets from Dublin and London.

[67] https://dev.twitter.com/streaming/overview
[68] http://boundingbox.klokantech.com/

Table A.1: Tweet datasets collected during the research for this thesis

City	BBox	Time Period	Tweets	Storage Volume
Dublin (IE)	-6.46,53.22,-6.03,53.45	26/11/2014 to 26/01/2016	6,359,500	11.6 GB
London (UK)	-0.51,51.31,0.30,51.71	26/11/2014 to 26/01/2016	36,513,800	66.8 GB
Campina Grande (BR)	-35.96,-7.31,-35.84,-7.18	23/01/2015 to 26/01/2016	14,970,300	24.2 GB
Maceió (BR)	-35.81,-9.71,-35.59,-9.39	26/01/2015 to 26/01/2016	15,505,425	25.5 GB
			73,349,025	128.1 GB

A.2 Filtering

The state-of-the-art approach to finding social media posts about a topic consists of using a manually curated set of keywords to filter the content stream (Bommannavar et al., 2016). Thus, a set of keywords in the domain of urban issues was built, including words such as pothole, graffiti, footpath, pathway, rubbish, litter, traffic, lighting, etc. This keywords set is composed of relevant keywords manually identified by looking into a number of FixMyStreet reports mapped in six urban issues categories: Graffiti; Leaks and Drainage; Litter and Illegal Dumping; Road/Path Defects; Street Lighting; and Tree/Grass Maintenance. Some keywords related to traffic issues were also included. In total, 39 keywords compose the keyword set for filtering the tweet datasets.

Seven rules compose the filtering strategy developed to be applied in the tweet datasets:

1. **Stemming keyword set:** instead of using a keyword on filtering tweets, the strategy adopts the word stem of the keyword. This ensures filtering not only by keyword but also including word variations (e.g. the stem "light" enables filtering by "light", "lights", "lighting", etc.). In order to process the stems of the keywords, the filtering task used the Porter Stemmer Online[69].

[69] http://9ol.es/porter_js_demo.html

2. **Language selection:** only English tweets and tweets shared by English native speakers were filtered.
3. **Short tweet removal:** tweets with length less than 20 characters were filtered out. Such character-length threshold was defined by identifying irrelevant messages in short tweets from the datasets.
4. **Similar removal:** using the Jaro-Winkler distance (Winkler, 1990) in the messages, tweets with distance (*similarity*) equals or greater than 0.75 were filtered out. This threshold was defined by testing different values and manually comparing the similarity among a number of messages using different thresholds.
5. **Replies removal:** tweets that are replies to other tweets were removed since they usually present lack of context and would not be useful in this study.
6. **Other sources removal:** tweets from other social media sources which might be irrelevant to the urban issues domain such as Foursquare check-ins and Instagram pictures were filtered out.
7. **User home location filled up:** only tweets with information about the user home location were kept in order to have tweet sample datasets with all location contexts available.

In total, 455 and 1,200 gold-standard tweet candidates from Dublin and London datasets, respectively, resulted from the initial datasets after filtering. These tweets compose the datasets to be manually labeled by the volunteers. A preliminary analysis on both tweet datasets shown that the filtering strategy worked well, as some urban issue reports could be found and the number of irrelevant tweets found was rare. Thus, the filtering process could ensure the occurrence of true positives on urban issues in the sample data to be manually labeled by the volunteers.

A.3 The Web Application for Labeling Tweets

In order to provide the human classification task over the gold-standard tweet candidates, a Web application for human-driven tweet labeling (the *Tweet Annotator*) was developed. The purpose of such application is to provide an easy GUI using a set of classes

for urban issues classification and Geographical Level of Details (GLoD) classification. Although Tweet Annotator has been initially developed to handle with classes in the domain of urban issues, the application can be easily adapted to other domains. The source code is available online[70]. Figure A.1 shows the interface of the Tweet Annotator presented to a volunteer during the classification task on a tweet.

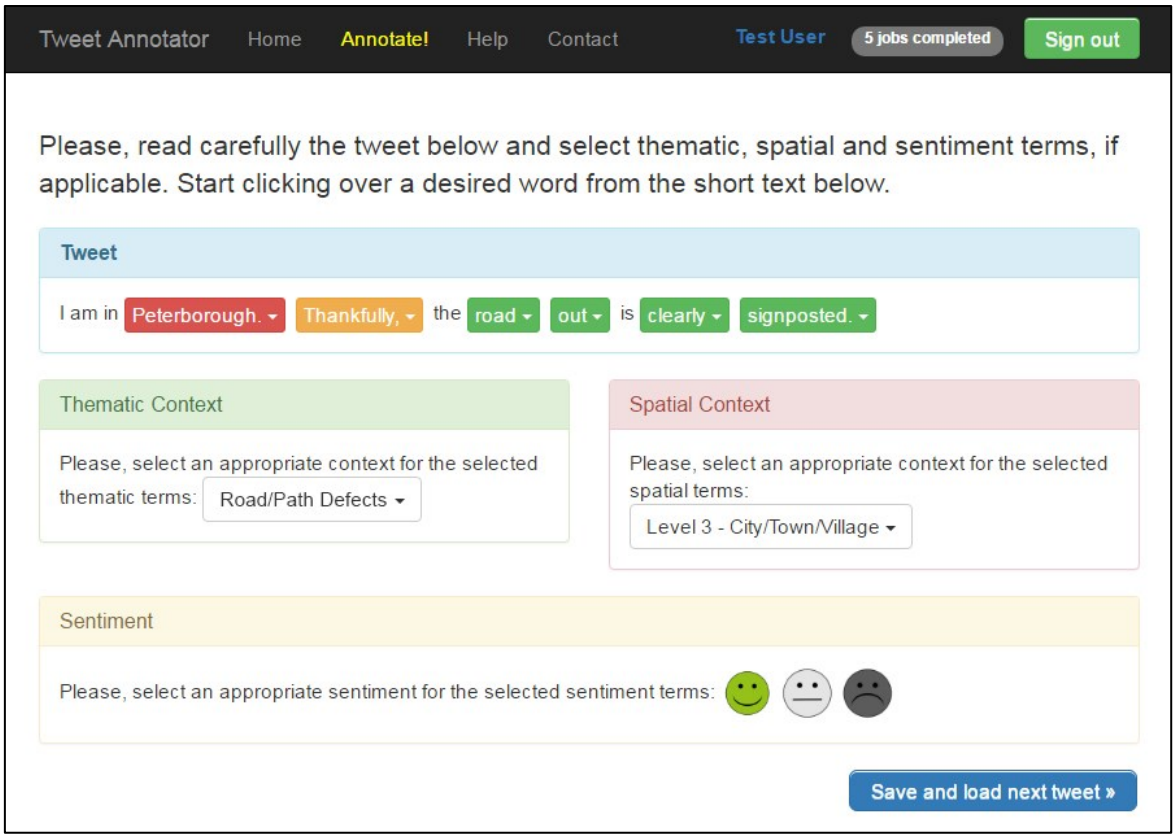


Figure A.1: An example of a labeling task on Tweet Annotator

In the Tweet Annotator, the labeling process starts with a human volunteer selecting a word from the tweet to be labeled (the blue box area shown in Figure A.1). The context to be attributed to a word should be selected by the volunteer when clicking on each word. There are three context options: thematic (green), spatial (red) and sentiment (orange). In the example shown in Figure A.1, a volunteer is assigning the thematic, spatial and semantic context to the tweet: the thematic class “Road/Path Defects”, related to the term “road out clearly signposted" from the tweet; the spatial class “Level 3 – City/Town/Village", related to the term “Peterborough” from the tweet; and the semantic class “Positive” (represented by the smile icon), related to the term “Thankfully”.

[70] https://github.com/maxmcz/tweet-annotator/

The Tweet Annotator provides 8 classes for classifying urban issues (thematic context), 7 classes for classifying mentioned geographical location according to their GLoD (spatial context), and 3 sentiment classes for classifying the sentiment polarity (sentiment context). The classes for the urban issues classification were defined based on the FixMyStreet categories. Thus, once a human volunteer identify a tweet that is related to an urban issue, such tweet is then classified in one of the following classes: 1) Graffiti/Flyposting; 2) Leaks and Drainage; 3) Litter and Illegal Dumping; 4) Road/Path Defects; 5) Street Lighting; 6) Traffic; 7) Tree/Grass Maintenance and 8) Other.

The classes for the GLoD were defined based on the extended GeoTree developed (see Figure 4.5 in Chapter 4). Such extended GeoTree includes GLoD's suitable for urban areas. Thus, once a geographical location mentioned in the tweet is identified by the volunteers, they must choose the GLoD according to the following classes: 1) Level 0 - Country; 2) Level 1 - Province; 3) Level 2 - County; 4) Level 3 - City/Town/Village; 5) Level 4 - District/Suburb; 6) Level 5 - Street/Road; and 7) Level 6 - Point of Interest.

Finally, the classes for the sentiment polarity come from related work on sentiment analysis such as Alves et al. (2015) and Wakamiya et al. (2015). Thus, the volunteers must choose the sentiment polarity according to the following classes: 1) Negative; 2) Neutral; and 3) Positive.

In case a tweet presents more than one thematic or spatial contexts, the volunteer must choose the most relevant thematic context according to the message and the higher GLoD for the spatial context. Thus, although a tweet may present several mentioned locations, the volunteer was instructed to label the most relevant one for the context of the tweet.

A.4 Manual Labeling

The set of recruited volunteers for the manual labeling tasks varied according to the tweet sample dataset. Ten volunteers (students and researchers from the UCD School of Computer Science, in Ireland) were recruited to perform the labeling on the gold-standard tweet candidates from Dublin tweet dataset through the Tweet Annotator. Fifteen volunteers (students and researchers from the UFCG Information Systems Laboratory, in

Brazil) were recruited to perform the labeling on the gold-standard tweet candidates from London tweet dataset.

The following criteria were established for the volunteers' recruitment to the Dublin tweet dataset labeling: they must be English-speakers and have been living in Dublin City for at least six months. For the volunteers' recruitment to the London tweet dataset labeling, the criteria were smoothed as it was not possible to work with London residents. Thus, such volunteers must speak Intermediate English at least. Moreover, the volunteers selected to perform the London tweet dataset labeling were trained regarding London geography, boroughs, street names and common slangs. They also had a look into some FixMyStreet reports from London in order to learn about the structured used by London residents to report urban issues.

The presented volunteers' recruitment criteria were defined in order to ensure a high Inter-Annotator Agreement (IAA) among the volunteers. Thus, the volunteers would be able to handle with both the usage of an informal spoken language and the usage of vernacular names on describing geographical locations, which requires a previous knowledge about the surroundings of the city. The volunteers were also trained to perform the labeling using the Tweet Annotator through an intuitive short demo video produced. In order to mitigate human errors, the volunteers had 30 business days to perform the task and they were instructed to not to label more than 20 tweets at a time.

The selected volunteers profile is the following: the average age is 28 years-old; regarding the gender, 32% are female and 68% are male; regarding the education level, around 30% are undergraduate students, 50% are masters or masters' students and 20% are PhDs or PhD students; and regarding the level of English, 48% are intermediate, 20% are advanced, and 32% are fluent.

The tweets were randomly sorted such that each tweet would be classified by four different volunteers at least. This method is widely used to ensure a good quality on the generated dataset. Only tweets which have the same contexts (or no context) assigned from at least 3 of the 4 volunteers was considered as labeled and then included in the final gold-standard labeled tweet datasets.

After selecting the tweets with agreement of 3 volunteers at least, all the tweets with geographical labeling were submitted to a manual geoparsing relying on Google

Maps. For each tweet, the GLoD assigned by the volunteers and the message words assigned to the spatial context were used to perform queries on Google Maps and retrieve the geographical coordinates to be assigned. Such process was partially-automated thanks to the Google Maps Geocoding API[71].

A.5 Corpora Overview and Statistics

In total, 1,494 tweets compose the corpora of labeled English tweets concerning urban issues, geographical locations and sentiment. 403 labeled tweets compose the corpus from Dublin (89% of IAA) whilst 1,091 labeled tweets compose the corpus rom London (91% of IAA).

In the Dublin corpus, 136 tweets (33.7%) were labeled as urban issues while 143 tweets (35.5%) were labeled as containing geographical locations mentioned. 75 tweets (18.6%) contains both thematic and spatial labels. 199 tweets (49.4%) have no context assigned. There is no sentiment context information from Dublin corpus since the labeling task performed with Dublin tweets used a preliminary Tweet Annotator version which did not provide support for sentiment labeling.

In the London corpus, 366 tweets (33.5%) were labeled as urban issues while 487 tweets (44.6%) were labeled as containing geographical locations mentioned. 254 tweets (23.3%) contains both thematic and spatial labels. 492 tweets (45.1%) have no context assigned. Regarding sentiment context, 216 tweets (19.8%) were labeled to negative polarity, 162 tweets (14.9%) were labeled to positive polarity and 713 tweets (65.4%) were labeled to neutral polarity. In addition, from the 366 tweets labeled as urban issues, 89 (24.3%) tweets present negative sentiment, 34 tweets (9.29%) present positive sentiment and 243 tweets (66.4%) present neutral sentiment. Such finding suggests that the automated identification of urban issues from social media should not rely on the tweet sentiment to perform such identification as the relationship "*sentiment* versus *urban issue*" seems to be weak. More discussion and details regarding the thematic and spatial contexts from the corpora produced in this research are provided in Section 5.1 (Chapter 5).

The volunteers were asked to complete a feedback survey containing concerning the labeling process performed aiming at identifying the difficulties faced by the volunteers

[71] https://developers.google.com/maps/documentation/geocoding/

when labeling the tweets. The main insights from the answers assigned by the majority of volunteers were:

1. The labeling task was considered OK to perform and the volunteers felt confident about their labels. However, the overall time spent during the task was reported as challenging;
2. The task of labeling the thematic context was considered OK, even though the volunteers have noticed missing classes such as "Cycling Problems" that seem to be very relevant for urban issues from Dublin City area;
3. The task of labeling the spatial context was considered easy. They considered the GLoD classes provided enough for the labeling task. Some volunteers highlighted that the geographical locations mentioned in the tweets was clear for them due to their previous knowledge regarding the city surroundings. Other volunteers highlighted the vagueness on Twitter users describing locations, suggesting other sources such as the tweet metadata, the user profile and the mentioned URLs would be needed in order to better identify such locations;
4. The task of labeling the sentiment context was considered OK. However, some volunteers reported some difficulties on labeling sentiment in tweets with irony/sarcasm;
5. The volunteers have considered the demo video and the information provided in the Tweet Annotator enough to perform a good job. In addition, the overall usability of the Web application was considered good;
6. Some labeling challenges were reported concerning tweets with unreadable characters or uncommon abbreviations, and tweets with more than one thematic or spatial context to be assigned.

Finally, the log data acquired by the Tweet Annotator during the labeling tasks performed on the tweets shown:

- The most costly tweet lasted 57 seconds to be annotated by a volunteer on labeling the Dublin dataset, while the most costly tweet in the London dataset lasted 60 seconds to be annotated;

- 19.5 seconds was the average time spent on labeling a tweet in the Dublin dataset (thematic and spatial contexts only), while 29.7 seconds was the such an average time in the London dataset (thematic, spatial and sentiment contexts);
- 817 distinct tokens from the tweet messages were selected by the volunteers to assign the thematic context in the tweets from the Dublin dataset, including hashtags, user mentions and stop words; while 1,652 distinct tokens were selected in the London dataset;
- 413 distinct tokens from the tweet messages were selected by the volunteers to assign the spatial context in the tweets from the Dublin dataset, including hashtags, user mentions and stop words; while 1,263 distinct tokens were selected in the London dataset;
- Finally, 946 distinct tokens from the tweet messages were selected by the volunteers to assign the sentiment context in the tweets from the London dataset, including hashtags, stop words and emoticons.

The generated gold-standard tweet corpora are available online[72] to be used in other researches.

[72] https://github.com/maxmcz/urban-issue-corpus/

Appendix B

Sample Tweet in JSON format

This appendix presents a JSON file containing a sample tweet which reports a real urban issue reported in Dublin, Ireland. Such tweet was harvested during the experiments through Twitter API, processed by Social2AGI and also manually annotated by human volunteers. Listing B.1 presents the raw JSON file as it was delivered from Twitter servers. For privacy reasons, some information which may identify the person behind the Twitter user was intentionally hidden.

Listing B.1: Sample tweet in JSON format


```
{
  "filter_level": "low",
  "retweeted": false,
  "in_reply_to_screen_name": null,
  "possibly_sensitive": false,
  "truncated": false,
  "lang": "en",
  "in_reply_to_status_id_str": null,
  "id": <HIDDEN>,
  "in_reply_to_user_id_str": null,
  "timestamp_ms": "1424972277307",
  "in_reply_to_status_id": null,
  "created_at": "Thu Feb 26 17:37:57 +0000 2015",
  "favorite_count": 0,
  "place": {
    "id": "7dde0febc9ef245b",
    "bounding_box": {
      "type": "Polygon",
      "coordinates": [
        [
          [ -6.3873911, 53.2987449 ],
          [ -6.3873911, 53.4110598 ],
```

```
          [ -6.1078047, 53.4110598 ],
          [ -6.1078047, 53.2987449 ]
        ]
      ]
    },
    "place_type": "city",
    "name": "Dublin City",
    "attributes": { },
    "country_code": "IE",
    "url": "https://api.twitter.com/1.1/geo/id/7dde0febc9ef245b.json",
    "country": "Ireland",
    "full_name": "Dublin City, Ireland"
  },
  "coordinates": {
    "type": "Point",
    "coordinates": [ -6.26313, 53.344301 ]
  },
  "text": "Dame St such badly laid out street. 4 traffic lanes, narrow
footpaths, dangerous for cyclists who make up significant share of road
users.",
  "contributors": null,
  "geo": {
    "type": "Point",
    "coordinates": [ 53.344301, -6.26313 ]
  },
  "entities": {
    "trends": [],
    "symbols": [],
    "urls": [],
    "hashtags": [],
    "user_mentions": []
  },
  "source": "Twitter for Android",
  "favorited": false,
  "in_reply_to_user_id": null,
  "retweet_count": 0,
  "id_str": <HIDDEN>,
  "user": {
    "location": "Cork & Dublin, Ireland",
    "default_profile": false,
    "profile_background_tile": true,
    "statuses_count": 19051,
    "lang": "en",
    "profile_link_color": "0084B4",
    "profile_banner_url": <HIDDEN>,
    "id": <HIDDEN>,
    "following": null,
    "protected": false,
    "favourites_count": 6783,
    "profile_text_color": "333333",
    "verified": false,
    "description": <HIDDEN>,
    "contributors_enabled": false,
    "profile_sidebar_border_color": "FFFFFF",
    "name": <HIDDEN>,
    "profile_background_color": "C0DEED",
    "created_at": "Mon Jan 11 21:28:17 +0000 2010",
    "default_profile_image": false,
    "followers_count": 1855,
```

```
    "profile_image_url_https": <HIDDEN>
    "geo_enabled": true,
    "profile_background_image_url": <HIDDEN>,
    "profile_background_image_url_https": <HIDDEN>,
    "follow_request_sent": null,
    "url": "http://instagram.com/<HIDDEN>",
    "utc_offset": 0,
    "time_zone": "Dublin",
    "notifications": null,
    "profile_use_background_image": true,
    "friends_count": 1622,
    "profile_sidebar_fill_color": "DDEEF6",
    "screen_name": <HIDDEN>,
    "id_str": <HIDDEN>,
    "profile_image_url": <HIDDEN>,
    "listed_count": 38,
    "is_translator": false
  }
}
```

Appendix C

Social2AGI: A system prototype

Chapter 4 presented the proposed approach to the automated identification of urban issues from social media and discussed some issues regarding preprocessing techniques, thematic and geographical analysis. This appendix presents a system prototype developed to perform such identification in practice and to evaluate the proposed ideas.

It is proposed the Social2AGI, a system which processes social media posts and produces AGI by considering geographic and thematic aspects related to a specific domain. The main idea of the Social2AGI is to identify sparse relevant information from social media data. Social2AGI is not an event detection/management system, since the purpose of the system is not to analyze trends or clusters of data in a space-time. Social2AGI is supposed to produce spatiotemporal data concerning a specific application domain which is not supported by event detection, such as urban issues.

Social2AGI implements the proposed approach described in details in Chapter 4. Social2AGI is implemented using Java, with a Web service which retrieves and returns external requests. Notice that the proposed system is not an online system which performs real time identification of urban issues. It is estimated a time window at around one hour between a tweet being harvested and fully processed, which may vary according to the amount of tweets per minute in each time.

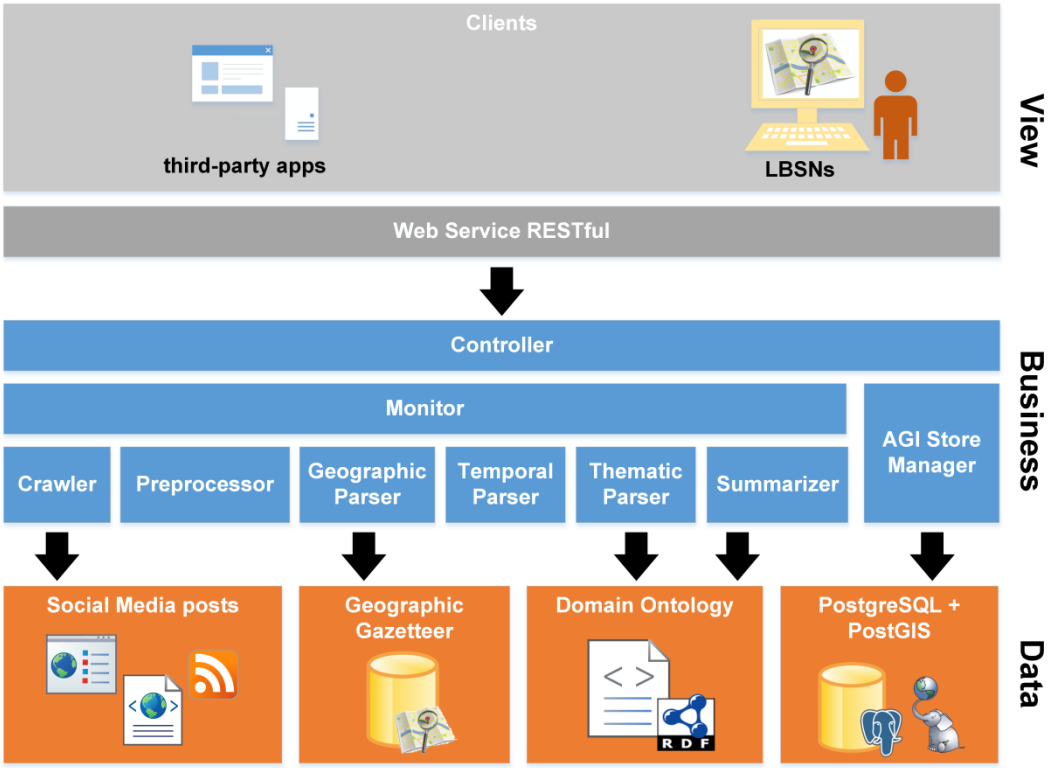


Figure C.1: Social2AGI System Architecture

The Social2AGI architecture is illustrated in Figure C.1. Social2AGI implementation has a multi-layer architecture which contains three main layers: view, business logic and data.

The business layer comprises eight interconnected modules. The "Controller" is responsible for handling requests from the Web service and works with the "AGI Store Manager" module, which provides access to the AGI data source. The "Monitor" is responsible for monitoring the harvesting and processing of social media posts. The "Crawler" is responsible for keeping the streaming which harvests social media posts in real time running. As social media posts are harvested and stored in the temporarily local data store, the "Monitor" sends such posts to the "Preprocessor", which is in charge of starting the processing of the posts. The remaining modules process the social media posts as detailed in the proposed approach presented in chapter 4.

The data layer contains four different data sources used by the Social2AGI. The first one stores social media posts which are further processed by Social2AGI. The second consists of a geographic gazetteer, a database of place names and their respective geometries, which is required by the geographic parser to resolve toponyms in the text. The

third data source is a domain ontology that provides the rules and vocabulary for the thematic parser. In this specific case, the UIDO ontology is used as a knowledge base in the Social2AGI. Finally, the fourth data source consists of a DBMS provided by PostgreSQL[73] and the spatial extension PostGIS[74], which stores the AGI produced.

The view layer is based on Web service requests/responses through a pre-defined command "request" and a list of search parameters. The view layer is supposed to be used by third-party or LBSN applications interested in retrieving, showing or analyzing the AGI shared by the Social2AGI. Table C.1 describes the list of parameters which can be used in a request command into the Social2AGI.

Table C.1: List of parameters for the Social2AGI request command

Command: http://__URL__/social2AGI/request?parameter...		
Parameter	**Value Type / Range**	**Description**
`text_contains`	keyword1, keyword2, …	To retrieve AGI records by searching for keywords in the text.
`max`	Integer	The maximum amount of results to be retrieved.
`start_date`	Date	To search AGI records from a given start date.
`end_date`	Date	To search AGI records until a given end date.
`bbox`	lat lng, lat lng, lat lng, lat lng	To perform geographical search of AGI records. The search is based on a bounding box area.

The response of a request command is in JSON format as the example shown in Listing 4.5 presented in Chapter 4.

[73] http://www.postgresql.org/
[74] http://postgis.net/

Appendix D

URL list of FixMyStreet RSS feeds

This appendix provides a list containing all the URLs of the FixMyStreet RSS feeds used by the developed crawler to harvest the FixMyStreet reports which comprise the FixMyStreet dataset described in Section 5.1. Listing D.1 lists all the URLs used to harvest Greater Dublin reports. Listing D.2 lists all the URLs used to harvest Greater London metropolitan reports.

Listing D.1: URL list for FixMyStreet reports from Dublin

```
http://fixmystreet.ie/rss/reports/Dublin+City
http://fixmystreet.ie/rss/reports/Dún+Laoghaire-Rathdown
http://fixmystreet.ie/rss/reports/Fingal
http://fixmystreet.ie/rss/reports/South+Dublin
```

Listing D.2: URL list for FixMyStreet reports from London

```
https://www.fixmystreet.com/rss/reports/Barking+and+Dagenham
https://www.fixmystreet.com/rss/reports/Bexley
https://www.fixmystreet.com/rss/reports/Brent
https://www.fixmystreet.com/rss/reports/Barnet
https://www.fixmystreet.com/rss/reports/Bromley
https://www.fixmystreet.com/rss/reports/Camden
https://www.fixmystreet.com/rss/reports/City+of+London+Corporation
```

https://www.fixmystreet.com/rss/reports/Croydon
https://www.fixmystreet.com/rss/reports/Ealing
https://www.fixmystreet.com/rss/reports/Enfield
https://www.fixmystreet.com/rss/reports/Hackney
https://www.fixmystreet.com/rss/reports/Hammersmith+and+Fulham
https://www.fixmystreet.com/rss/reports/Haringey
https://www.fixmystreet.com/rss/reports/Harrow
https://www.fixmystreet.com/rss/reports/Havering
https://www.fixmystreet.com/rss/reports/Hillingdon
https://www.fixmystreet.com/rss/reports/Hounslow
https://www.fixmystreet.com/rss/reports/Islington
https://www.fixmystreet.com/rss/reports/Lambeth
https://www.fixmystreet.com/rss/reports/Lewisham
https://www.fixmystreet.com/rss/reports/Merton
https://www.fixmystreet.com/rss/reports/Newham
https://www.fixmystreet.com/rss/reports/Redbridge
https://www.fixmystreet.com/rss/reports/Richmond+upon+Thames
https://www.fixmystreet.com/rss/reports/Southwark
https://www.fixmystreet.com/rss/reports/Sutton
https://www.fixmystreet.com/rss/reports/Tower+Hamlets
https://www.fixmystreet.com/rss/reports/Waltham+Forest
https://www.fixmystreet.com/rss/reports/Wandsworth
https://www.fixmystreet.com/rss/reports/Westminster
https://www.fixmystreet.com/rss/reports/Westminster/St+James's
https://www.fixmystreet.com/rss/reports/Royal+Borough+of+Greenwich
https://www.fixmystreet.com/rss/reports/Kensington+and+Chelsea
https://www.fixmystreet.com/rss/reports/Kingston+upon+Thames