FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO



Urban transport evaluation using knowledge extracted from social media

Francisco André Barreiros Murços

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

Supervisor: Tânia Daniela Lopes da Rocha Fontes

Co-Supervisor: Rosaldo José Fernandes Rossetti

September 17, 2021

© Francisco Murços, 2021

Abstract

Public opinion is nowadays a valuable data source for many sectors. Regarding the transportation and mobility sector, it is possible to collect information on real-time with reduced costs compared to other methods of information extraction.

In this dissertation, we defined a methodology to extract knowledge from messages collected from Twitter to analyse urban mobility. The methodology was structured according three main modules: system configuration, data analytics and visualization. The messages used for the demonstration of the proposed methodology were extracted during two months for three different cities: New York, London and Melbourne.

The text extraction from social media and its analysis are very time-consuming tasks due to the volume of the messages produced. Each message extracted from Twitter is, normally, short, informal and with a lot of slang or misspellings. To deal with that matter, by using NLTK (Natural Language Toolkit) tool, NLP (Natural Language Processing) techniques were applied so the text could be cleared and understandable by the algorithm.

For the classification of travel related messages, a BERT (Bidirectional Transformers for Language Understanding) embedding model was used. The model is pre-trained, unsupervised and was released in 2018. In order to understand if a simple model could have good performance, an *unigram* approach was used. Three lists of travel-related words were used: (i) a small list with 10 traveled-related words, (ii) a medium list with 35 traveled-related words and (iii) a big list with 344 traveled-related words. The results show a high model performance with precision and accuracy higher than 0.80 and 0.90, respectively. Popular words are *train*, *walk*, *street*, *car*, *station*, *street* and *avenue*. Consistent results were obtained for all the three cities assessed.

To evaluate the public opinion, the messages related to transportation and mobility were classified according to its sentiment. Then, to evaluate the polarity of the messages (positive, neutral or negative), VADER (Valence Aware Dictionary and sEntiment Reasone) sentiment tool was used. VADER is an easy tool to use and has great compatibility with social media messages and informal texts. It is a lexicon and rule based tool that calculates the compound value of text emotion according to its words. The developed methodology attained good performance results for the sentiment analysis where the average value of precision scored 0.77 while recall, accuracy and F1-score attained around 0.78. A specific analysis was made regarding a car crash event on New York on May 18, 2017. This analysis demonstrates that the methodology is capable of recognizing spacial changes and mobility flows directing to the potential causes of its origin.

The developed work allows the conclusion that the proposed methodology can be very helpful to transport engineers, urban planners, researchers and policymakers in getting insight into public opinions regarding urban mobility.

Resumo

Atualmente, as redes sociais constituem uma fonte de dados valiosa para vários setores de actividade. No sector da mobilidade, as redes sociais online permitem obter informação em tempo-real a um baixo custo, quando comparado com outros métodos de recolha de informação.

Nesta dissertação definiu-se uma metodologia para extrair conhecimento de mensagens extraídas do Twitter para analisar a mobilidade urbana. Esta metodologia foi estruturada em três módulos principais: configuração do sistema, análise de dados e visualização. As mensagens usadas para a demonstração da metodologia proposta foram recolhidas ao longo de dois meses para três cidades distintas: Nova Iorque, Londres e Melbourne.

A extração de textos das redes socias e a posterior análise são tarefas muito demoradas devido ao alto volume de dados produzido. Cada mensagem extraída do Twitter é, normalmente, curta, informal e com muita gíria ou erros gramaticais associados. Para tratar estas questões, recorrendo à ferramenta NLTK (Natural Language Toolkit), técnicas de NLP (Natural Language Processing) foram aplicadas para que o texto fosse limpo, adequado e compreendido pelo algoritmo.

Para a classificação das mensagens relacionadas com transportes, utilizou-se o modelo BERT (Bidirectional Transformers for Language Understanding) embedding. Trata-se de um modelo não-supervisionado pré-treinado lançado em 2018. No intuito de perceber se um modelo simples pode ter uma boa performance, utilizou-se uma abordagem *unigram*. Três listas de palavras relacionadas com transportes foram usadas: (i) uma lista pequena de 10 palavras, (ii) uma lista média com 35 palavras, e (iii) uma lista grande com 344 palavras. Os resultados da aplicação deste modelo monstram que este apresenta uma performance elevada, com a precisão e exatidão a registar valores superiores a 0.80 e 0.90, respetivamente. As palavras mais populares são *train*, *walk*, *street*, *car*, *station*, *street* e *avenue*. Os resultados obtidos são consistentes para as três cidades.

Para a avaliação da perceção da opinião pública, as mensagens relacionadas com tráfego foram classificadas quanto ao seu sentimento. Para avaliar a polaridade das mensagens (positivo, neutro ou negativo), utilizou-se a ferramenta VADER (Valence Aware Dictionary and sEntiment Reasone) sentiment. O VADER é uma ferramenta de fácil utilização e com boa compatibilidade com mensagens de redes sociais e textos informais. É baseada em campos lexicais e regras para calcular o valor composto do sentimento de um texto de acordo com as palavras usadas. A metodologia desenvolvida obteve bons resultados de performance na análise de sentimentos. O valor médio da precisão atingiu 0.77 e a exatidão atingiu 0.78. Foi feita uma análise a um evento específico que envolveu um acidente de carro em Nova Iorque no dia 18 de maio de 2017. A análise efetuada, em particular a este dia demonstra que a metodologia é capaz de identificar alterações espaciais e de fluxos de mobilidade indicando quais as potenciais causas da sua origem.

O trabalho desenvolvido permitiu concluir que a metodologia proposta pode ser bastante útil para auxiliar gestores de tráfego, planeadores urbanos, investigadores e formuladores de políticas a obter informações sobre as opiniões públicas sobre mobilidade urbana.

Acknowledgements

The process of working on my dissertation was one of the most challenging experiences in my life. There were some good moments but there were also some down ones when I needed help to get back on the right track.

Therefore, my great gratitude to all my friends that always believed in me and tried to pick me up when I was not at my greatest. Thank you for being there when needed and for playing an important role in my life.

To Professor Tânia Fontes and Professor Rosaldo Rossetti, thank you for pushing me to my limits with the ambitious project we developed and for all the support I was given. You made me want to grow and improve myself every time we met during the dissertation.

One of the most important people during the process of working on my thesis was my girlfriend, Marta Dias Ferreira. I want to thank you for all the motivation and love that you gave to me. You made me feel confident that I was worth to be lucky.

Thank to my parents, Maria Margarida Barreiros and Francisco Joaquim Murçós, and to my lovely sister, Marta João Murçós, for taking care of me and for showing me the guidelines to be a better person.

This work is financed by the ERDF - European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/ECITRA/32053/2017 - POCI-01-0145-FEDER-032053.

Without all of you, this wouldn't be possible. Thank you!

Francisco Murços

"Progress cannot be generated
when we are satisfied with existing situations."

Taiichi Ohno

Contents

1 Introduction **1**
- 1.1 Context . . . 1
- 1.2 Motivation . . . 2
- 1.3 Aims and Goals . . . 2
- 1.4 Document Structure . . . 3

2 Literature review **5**
- 2.1 Artificial Intelligence . . . 5
- 2.2 Word Embedding . . . 6
- 2.3 Text Analysis . . . 8
  - 2.3.1 Text Classification . . . 10
  - 2.3.2 Sentiment analysis . . . 11

3 Methodology **13**
- 3.1 Architecture . . . 13
- 3.2 Data . . . 15
- 3.3 Technologies and tools . . . 15
- 3.4 Pre-processing . . . 16
- 3.5 Analytic methods . . . 17
  - 3.5.1 Text classification . . . 17
  - 3.5.2 Sentiment analysis . . . 20
- 3.6 Performance . . . 20

4 Results and discussion **23**
- 4.1 Pre-processing . . . 23
- 4.2 Text classification . . . 24
  - 4.2.1 Performance . . . 24
  - 4.2.2 Temporal analysis . . . 25
  - 4.2.3 Spatial analysis . . . 30
- 4.3 Sentiment analysis . . . 32
  - 4.3.1 Performance . . . 32
  - 4.3.2 Temporal analysis . . . 32
  - 4.3.3 Spatial analysis . . . 34

5 Conclusions and future work **39**
- 5.1 Conclusions . . . 39
- 5.2 Publications . . . 40
- 5.3 Future work . . . 40

A **43**
A.1 Plots with the neutral messages for New York on May 17, 2017 43
A.2 Statistical data regarding transportation on London, Melbourne and New York from TomTom index 45

References **47**

List of Figures

3.1 Global architecture overview. 14
3.2 BERT input representation [1]. 17

4.1 Number of messages with travel-related terms from the medium list 3.3 for London (a), Melbourne (b) and New York (c). 26
4.2 Percentage of tweets per hour of the day of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i). 27
4.3 Average number of tweets per hour of the day of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i). 28
4.4 Average number of tweets per day of the week of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i). 29
4.5 Geographical distributions of tweets for London (a), Melbourne (c) and New York (e) and geographical distributions of travel related tweets for London (b), Melbourne (d) and New York (f) according to the medium list of target words 3.3. . . 31
4.6 Average number of tweets per hour of the day considering the sentiment for London (a), Melbourne (b) and New York (c), according to the medium list of target words 3.3. 33
4.7 Average number of tweets per day of the week considering the sentiment for London (a), Melbourne (b) and New York (c), according to the medium list of target words 3.3. 34
4.8 Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment from 6a.m. to 12a.m., according to the medium list of target words (Table 3.3). 35
4.9 2017 Times Square car crash (Manhattan, New York) [2]. 36
4.10 Density and sentiment evolution of the messages one hour before and three hours after a car was crashed in Times Square, New York City, United States, at 11:53 EDT of May 18, 2017. 37
4.11 Number of messages with travel-related terms from the medium list 3.3 for New York one hour before the accident and three hours after. 38

A.1 Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment (neutral) from 6 a.m. to 12 p.m., according to the medium list of target words (Table 3.3). 43

A.2 Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment (neutral) from 12 p.m. to 12 a.m., according to the medium list of target words (Table 3.3). 44
A.3 Average traffic congestion in 2017 for London (a), Melbourne (b) and New York (c) according TomTom index. 45
A.4 Hourly average traffic congestion along the week of 14/09/2021 for London (a), Melbourne (b) and New York (c) according TomTom index. 46

List of Tables

2.1 Literature Review of Text Classification and Sentiment Analysis Studies Applied in the Domain of Transportation. 9

3.1 Characteristics of cities analysed. 15
3.2 Small list of target words related with transportation. 18
3.3 Medium list of target words related with transportation. 18
3.4 Big list of target words related with transportation [3] 19

4.1 Data characterization. 23
4.2 Performance results of text classification model using three lists of travel-related terms for New York, London and Melbourne. 24
4.3 Indicators. 25
4.4 Performance results of sentiment analysis using the medium list of travel-related terms 3.3 for London, Melbourne and New York. 32

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Transformers for Language Understanding
BoE	Bag-of-Embeddings
BoW	Bag-of-Words
CAVs	Connected and Autonomous Vehicles
DT	Decision Trees
GPS	Global Positioning Systems
IoT	Internet of Things
ITS	Intelligent Transportation Systems
JSON	JavaScript Object Notation
LDA	Latent Dirichlet Allocation
LIWC	Linguistic Inquiry and Word Count
LR	Linear Regression
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
RF	Random Forest
SM	Social Media
SVM	Support-Vector Machine
VADER	Valence Aware Dictionary and sEntiment Reasone

Chapter 1

Introduction

This introductory section presents the theme contextualization 1.1 where the need of mobility and the present methods of evaluating the quality of the transport services are introduced. In the motivation section 1.2, the reason to explore a new approach, in this case, the Social Media (SM) knowledge extraction is explained. At last, the goals section 1.3 describes the main targets of this thesis and section 1.4 refers to how the document is structured.

1.1 Context

In 2018, 55% of the world's population is residing in urban areas, and by 2050, 68% of the world's population is projected to be urban [4]. With the general growth of cities, the need of mobility increases. As result, the levels of traffic congestion have been raising dramatically worldwide. Passenger traffic is projected to increase, with respect to 2005, by around 34% in 2030 and 51% by 2050 [5].

Congestion costs Europe about 1% of gross domestic product (GDP) each year [6]. According to the White Paper on transport (2011) the congestion costs will increase by about 50% by 2050. As consequence, the social costs of accidents and noise would continue to increase. This document shows that more resource-efficient vehicles and cleaner fuels are not the only solution to the problem of congestion as they need to be accompanied by an efficient core network for multimodal intercity travel and transport. Therefore, a long-term strategy related to the transformation of the transportation system in Europe is needed in order to have a competitive and sustainable transportation system.

Infrastructures and transportation systems are usually projected by using origin-destination matrix defined using traffic surveys [7]. This is a time consuming method and rather limited as is supported by a reduced sample.

Other method are the automatic traffic counters usually available in cities. A traffic count is a count of traffic along a particular road, either done electronically or by people counting by the side of the road [8]. This method relies on a device which requires installation in a specific location.

Although the great temporal details, there are some limitations such as a spacial limitation is present in traffic counters method.

In an era of technology, the use of Social Media (SM) is quickly raising. This new technology allows any citizen to share information about any subject in a real-time which means that SM emerge as a potential source of traffic information. Using information from the SM, it would be possible to have a more spatial-time detailed approach regarding the traffic by reporting in real-time the actual needs of users based on emotion and sentiment analysis.

Public transport operators and city planners have been using SM not only to provide but also to collect information from their travellers. The information collected in this way can be used to assess the quality of the services provided. However, due the nature of the information (mainly based on personal opinions), such data has not been used properly.

1.2 Motivation

Mobile devices and SM are the main agitators of big data statistics analysis. A staggering amount of data is produced daily by billions of users and thousands of businesses interactions. Therefore, to process all the data a great designed data mining and business intelligence are needed [9].

As previously mentioned on section 1.1, SM are rising very fast. Nearly 60% of the world's population (3.8 billions people) is already online, and the latest trends suggest that more than half of the world's total population is using SM [10]. Moreover, about half of the 3.7 hours that people spend using mobile phones each day is spent using social and communications apps [10]. So, people have easy access to a lot of information in real-time.

A staggering amount of data is produced daily by in SM by billions of users. The information available in SM covers lots of sector like health, sports, politics, and education. As result, SM are the main agitators of big data statistics analysis. Therefore, to process all the data a great designed data mining and business intelligence are needed [9].

In this dissertation, the mobility and transportation sector will be approached. The work will focus on how the information extracted from the SM may be a good way to get a city's traffic information.

1.3 Aims and Goals

The main objective of this project is to develop a methodology that allows to easily identify operational problems, like regular traffic congestion or delays in public transport, based on the opinions provided in social networks. The specific goals of this work are:

1. To define a methodology to evaluate the transports using the information provided through social media;

2. To evaluate if a simple methodology (*unigram* approach for text classification) is able to monitor urban mobility;

3. To demonstrate the applicability of such methodology on three case-studies: New York City, London, and Melbourne.

To achieve these goals, the project will use data mining tools as well as the definition of indicators, to understand the variability (in time and in space) of the quality of the transport services.

1.4 Document Structure

The remaining part of this document is structured as follows:

- Chapter 2 begins with a contextualization on transportation systems domains and explains a collection of related works using social media to extract knowledge;
- Chapter 3 presents the proposed architecture for extracting transportation knowledge from Twitter;
- Chapter 4 is where the main results are presented and explained;
- Chapter 5 is the last chapter of the document where conclusions and future work are exposed.

Chapter 2

Literature review

Intelligent Transportation Systems (ITS) were initially developed for the recognition of communications and computing's impact in the transportation field, exclusively in highway traffic [11]. Nowadays, ITS is a much wider concept that deals with information, not only from highway traffic, but also from public transportation such as trains, buses, or flights arrivals or even their routes [12].

The data produced in ITS is clearly measureless due to the great deployment over Global Positioning Systems (GPS), Connected and Autonomous Vehicles (CAVs), and other sensors and sources [11]. This amount of data allows deep knowledge regarding transportation systems, so methods of Smart Mobility [1] can improve, or even be created.

In addition, nowadays the Internet of Things (IoT) and Artificial Intelligence (AI) are upgrading the ITS concept in different transport modes including air, rail and sea, and are trying to minimize the main issues regarding this matter such as traffic jams or delays [14]. Once crowdsensing is increasing with the raise of Social Media, it is possible to arrange new, modern, and hopefully better methods to study and analyse what people feel and have to say about transportation issues.

In this chapter some essential and helpful concepts to understand how Social Media can be explored towards the main goal of improving the quality of transports will be analysed. First, the Artificial Intelligence (AI) field is explained (section 2.1), focusing on Natural Language Processing (NLP) and Machine Learning (ML). Next, different learning approaches for the extraction and analysis of Social Media data are presented (section 2.2). Although texts can be ambiguous and unorganized, after a pre-processing stage they can be evaluated according to emotions. To this end, Text Analysis (section 2.3) techniques applied to Social Media are analysed. In this section the analysis is focused on Text Classification (sub-section 2.3.1) analysis and Sentiment Analysis (sub-section 2.3.2) techniques applied to a transport context.

2.1 Artificial Intelligence

Last decade, Artificial Intelligence (AI) systems has achieved great performances in many tasks that were previously considered as unattainable. The rise of available information and hardware

[1] Smart Mobility is a concept related to how people get around in a clearer, safer, and more efficient way [13].

improvements combined with new and optimized algorithms provided advances in the field. These advances can be attributed to high-quality open-source libraries that allowed researchers and developers to code and test models [15].

AI is a representation of human-like intelligence by a machine that mimics the capabilities of the human mind. A process of learning is required from examples, experiences, language understanding and responding, making decisions, among others. With this process, a machine may be capable of performing functions a human might perform [16]. However, AI is nowadays called as weak AI once it is designed to perform special tasks, as the previously mentioned, outperforming humans. If a general AI was designed, it would outperform humans at almost every task [17].

Two of the subfields within artificial intelligence are Natural Language Processing (NLP) and Machine Learning (ML).

NLP involves extraction and analysis of written language [18, 19]. It supports text analytics by facilitating the understanding of sentence structure and meaning through statistical and ML methods. In other words, NLP is a way for computers to analyze and understand the meaning of human language in a smart and useful way [17]. Syntactic and semantic analysis are the main NLP tasks that analyse words interactions and the structure of the sentences. Both tasks have subtasks essential to the NLP algorithm such as tokenization, part-of-speech tagging, stemming and *stopword* removal, text classification, and sentiment analysis. There are many sources of unstructured data in the form of text and SM are one of them. Extracting insights form text can be a hard and time-consuming task due to its unstructured nature and sometimes ambiguous, but NLP techniques can help to automatically structure and analyse it quickly [20].

ML is present in information technology, statistics, probability, AI, psychology, neurobiology and many other disciplines. ML models are good representations of specific datasets and are used to solve several problems. It has become an advanced area by teaching the computers to simulate the human brain [21]. ML uses several sets of statistical techniques for problem solving. The algorithms used allow computers to learn how to perform some tasks from pre-processed data which is different of the common explicit programming. The ability of learning is one of the most important features of ML and is one of the most important steps of empowering the machines in order to help peoples' functions. The progress made in the field of ML is very important to NLP once makes it easier and quicker to preform [18, 22].

2.2 Word Embedding

Word Embedding is a representation model of vocabulary/words and is capable of capturing context of a word in documents, semantic and syntactic similarities, allowing words with similar meanings to have similar representations. In other words, word embeddings are vector representations of a particular word. So, they play an important role in NLP challenges and problems [23, 24, 25].

To represent a word-vector for a corpus data, a training process is required to the word embedding model using a large amount of data. For this purpose, ML has Supervised Learning methods and Unsupervised Learning methods that are frequently discussed together.

Supervised Learning consists on a ML task of learning a function that maps an input to an output based on labelled training data [26]. Each set of labelled training data is a group (or a vector) of pairs of input objects and desired output values. A Supervised Learning algorithm produces an inferred function based on the training data analysis, which serves the purpose of correctly mapping new data. The best scenario is that the inferred function is able to correctly determine the class labels for unseen instances [27]. A large number of Supervised Learning algorithms have been introduced last years [21], such as, Support-Vector Machine (SVM), Linear Regression (LR), Naïve Bayes (NB), Decision Trees (DT) or Random Forest (RF).

Unlike Supervised Learning algorithms, Unsupervised Learning doesn't need labeled data to either predict future outcomes or assign values to specific categories. Although Supervised Learning tends to be more accurate, it requires upfront human intervention to label the data [28]. Instead, Unsupervised Learning algorithms work on their own to find patterns and information that was previously not detected on unlabelled datasets. Therefore, Unsupervised Learning performs more complex processing tasks without the need of human intervention [29].

For word embeddings, there are several models such as as word2vec, doc2vec, GloVe [30] and BERT (Bidirectional Transformers for Language Understanding) [31]. The training process of word embeddings might be supervised with a neural network model, such as document classification, or unsupervised through statistics related to a document [23]. However, there are some limitations related to word embedding models, i.e., small amounts of data are not likeable to have good results by using a model with high dimensionality (which is the case of word embeddings), so this technology is great at dealing with Big Data and many researchers use it in a great variety of fields like industry, health, and transportation. In addition, sentiment information is neglected in word embeddings [30].

The unsupervised pre-trained BERT model was released in 2018. The developers of BERT define the model as capable of pre-training deep bidirectional representations from unlabelled text [1]. This model has been used in several NLP applications. The model is able to extract high quality language features from text, but the model can also be used fine-tuned on specific tasks, such as, entity recognition or question answering [32].

BERT embeddings have been used for information retrieval, keyword expansion and semantic search. These embeddings generate word and sentence embedding vectors which are high-quality feature inputs to downstream models. The translation of text to numerical representations is required by NLP models, which have been represented as indexed values (one-hot encoding) in the past for models like Word2Vec or Fasttext. The advantage over these models that BERT offers is that it produces dynamic word representations that depend on the context (nearby words) [33].

2.3 Text Analysis

Text Analysis is a ML technique that allows to automatically understand text data, such as tweets, emails, support tickets, product reviews, and survey responses, and can be used to extract specific information from thousands of unstructured datasets, or categorize by sentiment and topic. By delivering qualitative results, Text Analysis identifies important information within the text which is a challenge due to the ambiguity of human language [34].

There are some Text Analysis techniques, like Text Classification, Text Extraction, Word Frequency, or Clustering, each used for different purposes. There are some tasks within each technique that are useful to break down and understand text as a human would such as Sentiment Analysis tasks within Text Classification. These methodologies are widely used and have been used in several researches regarding transportation [12, 20].

Text classification is a process that assigns categories or tags to texts considering its content. Text classifiers are used to categorize, organize and structure almost any kind of text form turning it into a specific format value. This value can be expressed as numbers (sentiment analysis tasks), labels (multi-labelling tasks) or classes (binary or multi-class tasks) playing an important role in NLP.

Sentiment Analysis is a process used to detect negative or positive sentiment in texts. It is applied in many domains, such as, businesses, researches, governments, and organizations [35] to detect sentiment in social data and understand customers since they express their thoughts and feelings openly. The analysis can focus on polarity (*positive, negative, neutral*), feelings and emotions (*angry, happy, sad*), urgency (*urgent, not urgent*) or intentions (*interested, not interested*). It presents some benefits such as sorting data at scale, real-time analysis and consistent criteria [36].

In order to understand how Text Classification and Sentiment Analysis algorithms have been applied for extracting information of transport from SM, a literature review was conducted. Table 2.1 show a synthesis of that review.

Table 2.1: Literature Review of Text Classification and Sentiment Analysis Studies Applied in the Domain of Transportation.

Ref.	Local	Source	Data			Text Classification				Sentiment Analysis		Main goal
			Period	Year	N	Approach	Assumption	N.º of keywords	Data manually labelled	Method	Polarity	
[37]	Indonesia	T	-	-	1.2K	SVM	BoW	40	-	Linear SVM	P, N	Assessment of use of city public transportation
[38]	Portugal	T	2 months	2010	0.6M	SVM	Bootstrapping	-	4000	-	-	Real-time sensing of traffic information
[39]	U.S.	T	4 months	2010/2014	64K	Machine learning	-	Lexicon[d]	-	-	-	Evaluate users' opinions about public transit
[40]	Milan	T	2 years	2013	45K	SVM	-	∼50	1332	SVM	P, N, E	Evaluate the quality of service.
[41]	Santiago do Chile	T	3 years	2014/2016	110K	LDA	-	118	-	SentiStrength	P, N, E	Level of satisfaction of bus users
[42]	Melbourne	T	23 weeks	2015/16	293 K	-	-	13	-	-	-	Travel model choice assessment
[30]	New York, London	T, F, O[b c]	2 years	2016/18	-	LDA	word2vec, doc2vec	500	-	Lexicon2vec	P, N, E	Generates topics and features of transportation
[43]	Angkot	T	2 months	2016	272	SVM, DT, NB	-	40	7%	SVM, DT, NB	P, N	Assessment of customer satisfaction of online transport services
	Kopaja				184							
	Metro Mini				264							
	Transjakarta				481							
[44]	Salt Lake City	T	1 week	2017	403	LDA	BoW	15	-	Rsentiment	P, N	Assessment of transit rider' opinion about quality of service
[45]	Portugal	T	3 days	-	-	Supervised classification	BoW	1	-	Lexicon-based approaches	P, N, E	Assessment of how an event influences the volume of tweets about Uber and user perception
[12]	London	T	3 months	2017	5,6M	SVM, LR, RF	BoW, BoE, BoW + BoE	10	2000	-	-	Semantic analysis for Smart Cities
	New York				8.5M							
	Melbourne				0.8M							
	São Paulo				9.8M							
	Rio de Janeiro				18.8M							
[46]	Miami-Dade County	T	1 year	2017/18	0.4M	SVM, DT, NB	-	72	2000	Average Sentiment Value	0-1[a]	Evaluating the public opinion on transportation services
[47]	World	T, F, W, A, R	-	-	-	Protégé OWL	-	500	-	SentiWordNet	0-1[a]	Safe travelling
[48]	London	T, Q, I, K	-	-	-	LDA	-	-	-	-	-	Analyse public transit flow and detect major transit hubs

Sources: A: announcements, I: Instagram, F: Facebook, W: news, Q: Foursquare, R: traveling sites (web crawler), T: Twitter, K: Flicker, O: Others.
Classification approach: SVM: Support Vector Machine; DT: Decision Tree; NB: Naïve Bayes; LR: Logistic Regression; LDA: Latent Dirichlet Allocation; RF: Random Forest.
Classification assumption: BoW: Bag of Words; BoE: Bag of Embeddings.
Polarity: P: Positive, N: Negative, E: Neutral.
[a]higher value indicates positive attitudes; [b]New York Times; [c]TripAdvisor; [d]http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

2.3.1 Text Classification

Several text mining classification studies related with transportation have been conducted in the United States [39], particularly in New York [12, 30] and Miami [46]. Other cities worldwide, such as London [12, 30], São Paulo [12], Rio de Janeiro [12], Milan [40], Melbourne [12, 42], and Santiago do Chile [41] have explored data from the social media.

Twitter is the most popular data source used to conduct transport related studies (Table 2.1). A few contributions have focused on the analysis in this domain using data from other social networks, such as Facebook [30, 47], Foursquare [48], Instagram [48], and Flicker [48]. Some of the studies using SM data are focused only on geo-referenced messages. However, as in some social media as Twitter, the geo-referencing of tweets is not a default option. Thus, keywords associated with the name of transport system, codes of the bus stops or the number of the bus services have been used to better identify transport services [12].

Several lists of travel-related terms have been used in the literature for text mining classification. The number of words used in these studies is quite distinct from each study, ranging from 1 [45] to 500 [30, 47]. Qi et al. [46] define a list of 72 keywords for seven different transportation modes. The most popular travel-related terms in these studies are associated with the identification of transport mode (e.g., bike, bus, car, taxi, train) [42, 44, 46], and some of these lists were validated with transportation experts. In the domain of public transport, Haghighi et al., [44] and Anastasia et al., [43] use a list of 15 and 40 keywords respectively for public transit riders' satisfaction. Collins et al. [49] collected tweets containing keywords of train names in the city of Chicago. Effendy et al. [37] use a list of 40 Indonesian words by type of public transport. To infer the user satisfaction of public transport Mendéz et al. [41] defined a list of keywords for assessing topics related to the use of stops (frequency, app, information, stop, out of order, driver, overcrowding, social cleaniness, and regularity) and topics related to the use of buses (frequency, stop, information, out of order, App, and social).

In the domain of transport, supervised and unsupervised machine learning techniques have been used for text mining classification (Table 2.1). Text classification uses keyword search to extract transit-related messages. Popular machine learning supervised techniques are Support-Vector Machines (SVM) [12, 37, 38, 40, 43, 46], Decision Trees (DT) [43, 46], Naïve Bayes (NB) [43, 46], Latent Dirichlet Allocation (LDA) [30, 41, 44, 48], Logistic Regression (LR) [12] and Random Forest (RF) [12]. When it comes to the assumptions, the most common are Bag-of-Words (BoW) [12, 37, 44, 45] and Bag-of-Embeddings (BoE) [44, 45], with very few incidence on unsupervised techniques (word2vec and doc2vec [30]).

To compute LDA, the "topicmodels" package in R [44] and the Language Toolkit (MALLET) LDA are popular options. Steiger et al. [48] used an LDA model and then applied density-based spatial clustering (DBSCAN) to find clusters with points closely packed together. Farman et al. [30] propose an ontology, and LDA-based topic modelling called topic2vec, that extracts the most appropriate topics and features of document and neglects irrelevant words to enhance the document representation. Such ontology represents semantic knowledge that enriches an LDA

model to extract more accurate features from transportation texts [30].

2.3.2 Sentiment analysis

The Sentiment Analysis of Twitter data has gained much attention as a topic of research [50]. The ability to obtain information about a public opinion by analyzing Twitter data and automatically classifying their sentiment polarity has attracted researchers because of the concise language used in tweets [50].In the domain of urban planning and transportation by Mohammed Abdul-Rahman et al. (2021) built a framework to pre-process location-based social media big data [51].

Messages can be analyzed using ML methods [37, 40, 43] and lexicon-based methods [30, 45]. Examples of ML are the support vector machine (SVM) [37, 40, 43], Naïve Bayes (NB) [43], decision trees [43]. While ML methods need a training dataset to learn the model from corpus data, and a testing dataset to verify the model built, lexicon based method is based on a dictionary of words and phrases with a classification value (eg. positive and negative).

In the domain of transport, some studies labelling text as Positive or Negative [37, 43, 44], Positive, Negative or Neutral [40, 41, 42, 45], and zero or one [46, 47]. The approach of using supervised learning techniques to evaluate the sentiment on text is not as detailed and high performance as a lexicon and rule based technique. There are some Sentiment Analysis works [30, 41, 44, 45, 47] that are better at results.

In the field of sentiment analysis, the most widely used lexicon is SentiWordNet [37, 30]. However, when we have different senses of words, this approach may not obtain good results [30].

Currently, many tools, such as Linguistic Inquiry and Word Count (LIWC), used by Kahn et al. [52], offer the means of extracting advanced features from texts. However, most of these tools require some major programming knowledge and are difficult to interpret. There are some recent Sentiment Analysis tools are the Valence Aware Dictionary and sEntiment Reasoner (VADER) [53], TextBlob [53], and Afinn [54].

VADER can be used to determine the polarity of tweets and to classify them according to multiclass sentiment analysis. This tool is simple to use requiring simple programming code. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media once it works well with slang [53].

TextBlob performs more tasks than Sentiment Analysis once it is a Python library for processing textual data. It provides a simple API for diving into common NLP tasks such as part-of-speech tagging, noun phrase extraction or translation [53]. Regarding Sentiment Analysis, TextBlob is used to assign both polarity and subjectivity scores to the data [54] and performs strongly with more formal language usage.

Afinn is a lexicon that works on python and contains a dictionary with 3,300 words and their respective sentiment scores [54]. It is one of the simplest and most popular lexicons for Sentiment Analyisis. Although the ease of use and the good average results in previous studies, it might be a little outdated comparing to VADER and TextBlob [55].

Chapter 3

Methodology

In this chapter the adopted methodology is described along with its details and specificities. First, the architecture is explained which is the set of different modules to be performed (section 3.1). Then, the data used for the analysis is described (section 3.2). Some technologies and tools for data storage and data visualizations are explained (section 3.3), and the NLP techniques are explained on the pre-processing section once data needs some cleaning for further analysis (section 3.4). The classification technique of travel related messages and its emotion is mentioned (section 3.5). It is important to understand how the analytic methods work and how different conditions influence the performance.

3.1 Architecture

The proposed methodology was created to analyse social media messages related to transportation and includes three main modules: 1. System Configuration, 2. Data Analytics, and 3. Visualization.

1. The system configuration includes the extraction of social media data, in this case, from Twitter, and its storage on a local database. Configurations also include a set of pre-processing text operations to prepare the messages for the analysis by removing everything that is valueless, and the target filtering. The pre-processing is performed by NLTK (Natural Language Toolkit) which is an NLP tool.

2. Data analytics represent all the text classification procedures from the methodology. The text classification procedures include the pre-trained word embedding model (BERT base model (uncased)) that is implemented to filter only the messages related to transportation based on pre-defined lists of target words. Then, some manual classification of samples is included so the system's performance can be evaluated. The sentiment analysis is the last procedure of the data analytics module. It is important to analyse the opinions and emotions present in the messages. For that, VADER sentiment was used. Manual classification of samples was also made for the sentiment analysis to evaluate its performance.

3. The third module consists on the aggregation and plotting of the results so they can be read and interpreted. The plotting was performed with Matplotlib python library and Microsoft Excel software. This module allow the user to visualize the knowledge extracted from social media messages using charts and maps.

Figure 3.1 shows an overview of the methodology followed.

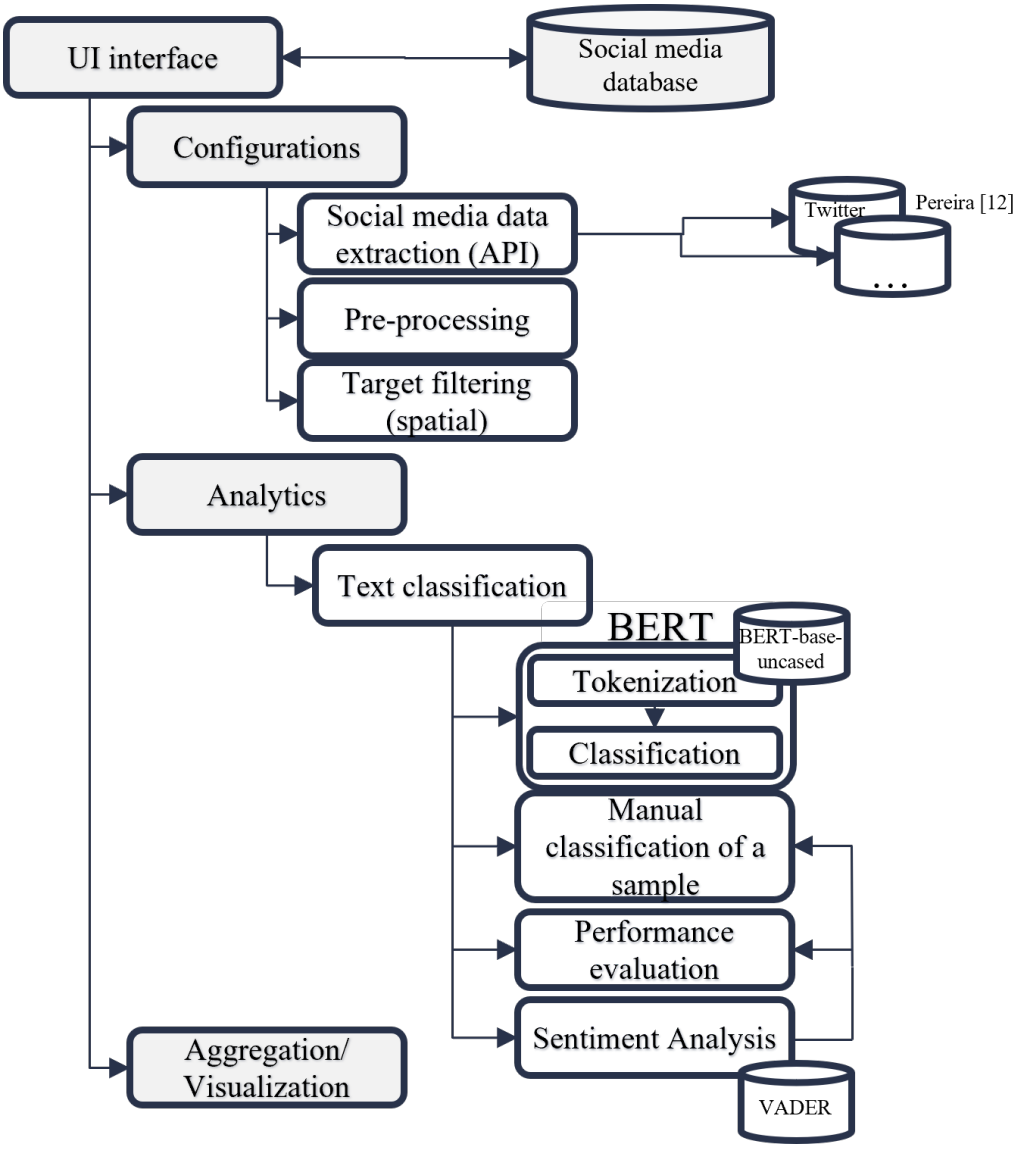


Figure 3.1: Global architecture overview.

3.2 Data

The data used were extracted from Twitter API. Twitter is a microblogging platform that allows users to share and access messages within a group of followers. A tweet is a message of up to 140 characters, images and links to other content. Tweets are classified in original tweets, replies, and retweets, and are in JSON format where each one is composed by several indicators. The most relevant are text content, geographic information and date and time.

Pereira [12] performed the process of extracting messages from Twitter. The messages collected are from the 16^{th} of May of 2017 to the 6^{th} of July of 2017. Data from New York, London and Melbourne were used in this study. All messages have a geographical reference: ones had specific coordinates associated and others had a bounding-box associated.

Among the three analysed cities, New York is the smallest one, in area, but has the highest number of residents. London has a population similar to New York, but has twice the area. Both in New York and in London the main transport is the metropolitan. On the other hand, Melbourne has about half of population of the other two analysed cities, and tram, train and bus are the popular means of transport. The main language used in the three cities is the English. Table 3.1 shows an overview of the main characteristics of the three cities analysed.

Table 3.1: Characteristics of cities analysed.

	New York	**London**	**Melbourne**
Area (km^2)	784	1 572	9 990
Number of residents	8.6 M	8.1 M	4.4 M
Main Transports	M, B, taxi	M, T, C, B	T, C, B
Climate	Humid subtropical	Temperate Oceanic	Temperate oceanic
Language	EN	EN	EN
Nº of messages extracted	9 151 579	5 815 757	835 737

Main transports: M: Metropolitan; T: tram; C: train; B: bus.

3.3 Technologies and tools

The system implemented follows the general platform proposed in [45]. Instead of a cradle, the Twitter Streaming API were used to extract the messages. The extracted data from Twitter is in the JSON format. Each message contains content, date and location. To read JSON files, json[4] Enconder and Decoder was used.

MongoDB[5], a NoSQL database (also known as non-relational database), was used to store the data collected, as well as the processed data. This database system was selected because it is compatible with the use of JSON files collected from Twitter. MongoDB is also compatible with Python programming language. In this dissertation, the database was design in a simple way where

[4] https://docs.python.org/3/library/json.html
[5] https://www.mongodb.com/

some main collections were defined, such as, *Total Data Collection*, *Pre-Processed Collection* and *Travel-Related Collection*. Some intermediate collections were used to improve the programming performance.

One of the most time consuming tasks of this methodology is the illustration of the results. Visualization of information collected from Twitter needs to be simple and objective so there are some data collection processes before the graphical plotting. For each city and each list of target words (Tables 3.2, 3.3 and 3.4), some analysis were made, such as, counting the travel related terms, counting the amount of travel related messages written on each hour of the day and on each day of the week, calculating the average amount of tweets on each day of the week or assemble the locations of each tweet. In order to plot the total number of tweets per hour of the day and per day of the week, CSV files were created and exported to Mircrosoft Excel and posteriorly plotted as bar charts. For the rest of the data, the plotting was made by using the Python library Matplotlib[6]. Matplotlib is a low level graph plotting library in Python that serves as a visualization utility. It is open source and can be used freely.

3.4 Pre-processing

Texts are represented by values stored in vectors and can be associated to very large lengths which may generate a problem. In order to avoid it, some adjustments may be made, such as, only considering words that appear at least a pre-defined number of times and ignoring some grammatical conjunctions [38].

The main objective of data pre-processing is to prepare the messages for text classification. For that, python's NLTK[2] (Natural Language Toolkit) is a good and easy to use NLP tool that was applied in this work. NLTK has incorporated most of the tasks like tokenization, stemming, lemmatization, punctuation, character count, and word count. Therefore, for each message a small and basic group of pre-processing operations were conducted, particularly:

- **Replace**: substitute integer occurrences for textual representations and contractions (e.g., 'it's' is replace by 'it is');
- **Cleaning**: removing hashtags, URLs, user mentions, non-ASCII characters (e.g. emoji's), retweets (RT), punctuation, and stopwords since these words do not add value to the text (e.g. in, a, an, the);
- **Lower-casing**: every word was converted into lowercase;
- **Lemmatization**: grouping together the inflected forms of a word (only performed for the verbs).

The BERT model is related to the WordPiece token vocabulary [56], so the stemming was not performed because it changes the root form of some words. For example, the word *happier* after

[6] https://matplotlib.org/
[2] http://www.nltk.org/

the stemming is changed to *happi* instead of *happy*. This event conflicts with BERT tokenization and embeddings.

3.5 Analytic methods

3.5.1 Text classification

An unsupervised pre-trained BERT model (Bidirectional Encoder Representations from Transformers) released in 2018 was used. The developers of BERT define the model as capable of pre-training deep bidirectional representations from unlabelled text [1]. This model has been used in several NLP applications. We used the model to extract high quality language features from text, but the model can also be used fine-tuned on specific tasks, such as, entity recognition or question answering [32].

BERT embeddings have been used for information retrieval, keyword expansion and semantic search. These embeddings generate word and sentence embedding vectors which are high-quality feature inputs to downstream models. The translation of text to numerical representations is required by NLP models, which have been represented as indexed values (one-hot encoding) in the past for models like Word2Vec or Fasttext. The advantage over these models that BERT offers is that it produces dynamic word representations that depend on the context (nearby words) [33]. The model was tested with three different lists of target words to evaluate its performance: a list with 10 travel-related words (Table 3.2), other with 35 travel-related words (Table 3.3) and a bigger one with 344 travel-related words (Table 3.4), based on Gal-Tzur et al. [3] that created a semi-automatically constructed transport lexicon.

The BERT architecture was built on top of Transformer with some variants. In this thesis, BERT Base (uncased) model is used. Such model is composed by 12 encoder layers, 12 attention heads and 110 million parameters. Every input embedding is formed by 3 different embeddings, as shown in the Figure 3.2: Position Embeddings (expresses the position of words in a sentence); Segment Embeddings (sentence pairs); Token Embeddings (specific token embeddings from the WordPiece token vocabulary) [57].

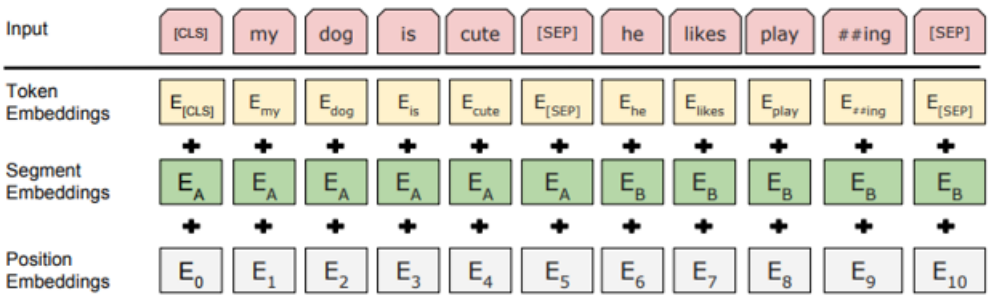


Figure 3.2: BERT input representation [1].

For the input text, BERT expects a tokenization including two special tokens: one that corresponds to the beginning of the sentence ([CLS]) and other that corresponds to the end of the sentence or the separation between two sentences ([SEP]). So, the tokenized text must have this format: "[CLS]" + *tokenized text* + "[SEP]".

In the fields of computational linguistics, *n-gram* is a sequence of *n* items from a given sample of text. The items can be characters, sequence of characters (e.g., words) or sequence of words. The length of the samples defines the *n* in the *n-gram*. In this work, each item corresponds to a word so an *unigram* approach were used.

In order to assess how the model performance is affected , three different list of traveled-related words were used:

- a small list with 10 traveled-related words focused on the name of the main urban transport modes (Table 3.2);
- a medium list with 35 traveled-related words (Table 3.3);
- a big list with 344 traveled-related words defined based on [3] (Table 3.4).

Table 3.2: Small list of target words related with transportation.

bicycle	bus	car	subway	train
bike	cab	metro	taxi	walk

Table 3.3: Medium list of target words related with transportation.

accident	boulevard	driver	motorcycles	taxi	uber
avenue	bus	drivers	road	taxis	underground
bicycle	buses	highway	station	traffic	van
bicycles	cab	metro	street	train	vans
bike	car	moto	streets	truck	walk
bikes	cars	motorcycle	subway	trucks	

Table 3.4: Big list of target words related with transportation [3]

accelerate	biker	congestion	driven	fueled	lane	nearest	pilot	route	stops	transferred	visitor
accelerated	bikes	connect	driver	fuels	license	network	pilots	routes	street	transferring	wait
acceleration	biking	connecting	drivers	gasoline	line	northbound	platform	routing	suburban	transfers	walk
accessibility	boarding	connection	drives	gates	lines	northern	platforms	running	subway	transit	walked
accidents	boards	connections	driving	gear	link	obstruction	points	runs	tanker	transmission	walkers
aerodrome	boat	continental	drop	going	localized	oil	priced	runway	tariff	transported	walking
airbus	booking	crossing	eastbound	heading	logistics	omnibus	promenade	runways	taxi	transport	walkway
aircraft	borders	crossings	eastern	helicopters	luggage	operates	propulsion	rush	taxis	transports	weekdays
airfield	brake	cycle	electrification	highway	mainline	overseas	queue	safely	terminal	travel	westbound
airline	brakes	cycling	emissions	highways	manual	parking	quicker	safer	ticket	travellers	wheeled
airlines	braking	cyclist	engine	hourly	meter	passenger	rail	safety	tickets	travelling	wheels
airport	breaks	cyclists	engines	hub	meters	passengers	railway	sailing	time	travelled	worldwide
airports	budgets	daytime	exhaust	incident	metro	passing	rd	schedule	timely	traveller	zone
airspace	bus	delay	exit	incidents	metropolitan	passport	reaches	seat	times	travels	zones
airways	buses	delayed	exploring	inland	mile	path	reliable	seating	timetable	trip	
alert	cab	delays	fare	interchange	miles	paths	rental	seats	tour	trips	
arrival	capacity	delivering	fares	intercity	minutes	pathway	return	services	tourists	truck	
arrivals	car	delivers	faster	internationally	mobile	pathways	returning	shuttle	tow	trucks	
arrive	cargo	departing	fleet	intersection	mobility	pavement	returns	sidewalks	towing	trunk	
arriving	carriage	departure	flights	intersections	monitor	payload	ride	sign	track	tunnels	
auto	carriers	depot	flooding	journey	motion	pedal	rider	southbound	tracking	turn	
automated	carrying	destination	flows	journeys	motor	pedestrian	riders	southern	tracks	turning	
automobile	cars	destinations	flying	junction	motorcycle	pedestrians	riding	speed	traffic	tyres	
automobiles	catering	direction	foreign	junctions	motorcycles	periodic	road	speeding	trail	taxi	
automotive	changing	directional	freeway	kilometer	motorized	periodically	roads	speeds	trailers	vacant	
aviation	channel	disruption	freight	kilometers	motorway	periods	roadside	station	train	van	
baggage	coaches	distance	frequencies	km	mounted	petrol	roll	stationary	trains	vans	
bicycle	collision	districts	frequent	kilometre	mounting	petroleum	rolling	stations	tram	vehicles	
bicycles	combustion	dock	frequently	kilometres	municipal	pickup	rolls	steer	trams	vehicle	
bike	commuter	drive	fuel	landing	navigation	pier	roundabout	steering	transfer	visiting	

3.5.2 Sentiment analysis

Sentiment analysis techniques are growing and gaining a lot of importance when it comes to SM analysis. In this work, messages were extracted from Twitter which is a great source of information about public opinion and emotion [43]. The sentiment analysis was performed by using a lexicon and rule-based tool called VADER[3] (Valence Aware Dictionary and sEntiment Reasone) [53] that is optimized for SM data. This tool is very simple to use, with easy programming code in Python, and is based on a list of lexical features which have a sentiment value assigned. Then, according to their semantic orientation, VADER calculates the text sentiment [58].

VADER sentiment returns the probability of an input text to be positive, negative or neutral. The sum of the three probabilities must be equal to 1. The compound score is also returned by VADER sentiment and it consists in a metric that calculates the sum of all the lexicon ratings which have been normalized between -1 and +1 (where the maximum value means the most positive sentiment). So, in order to label a message as positive, neutral or negative, it is important to analyse the compound value:

- a positive sentiment represents a compound value equal or higher than 0.05;
- a negative sentiment represents a compound value equal or lower than -0.05;
- a neutral sentiment represents a compound value between -0.05 and 0.05.

3.6 Performance

The performance of the classifier was defined by how well the result agrees with human judgement. The evaluation of the classifier was performed to text classification, as well as for the sentiment analysis results where, for each city and each list of target words, 1000 tweets were selected: 500 travel-related tweets and 500 non-related for the text classification performance; 500 positive tweets and 500 negative tweets for the sentiment analysis performance. In the case of sentiment analysis, neutral tweets were excluded from the performance evaluation as they are considered *non-labelled* because of the compound value calculated by VADER sentiment (between -0.05 and 0.05 where the compound value range is between -1 and 1). The selection was completely random and was set so it could find tweets created at different times. This set of messages were manually labelled in order to confirm whether they were travel-related or not, and with positive sentiment or not. Two human annotators performed this classification. No specific training has been provided to the human "labellers" since the manual labelling process was subjective.

Accuracy [37, 43, 46, 47], precision [12, 43, 46, 47, 38], and recall [12, 43, 46, 47, 38] are the most popular performance measures. However, the F1-score [12, 46], the Root Mean Square Error (MSE) [47], and Mean Average Error (MAE) [47] have been also measures commonly applied. Chaturvedi et al., [59] validate the results obtained using national and local newspapers. In our

[3] https://pypi.org/project/vaderSentiment/

study the performance of the classifier was done using accuracy 3.1 and a group of three per-class metrics: precision 3.2, recall 3.3, and the F1-score 3.4.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn} \tag{3.1}$$

$$Precision = \frac{tp}{tp+fp} \tag{3.2}$$

$$Recall = \frac{tp}{tp+fn} \tag{3.3}$$

$$F1-score = \frac{Precision*Recall}{Precision+Recall} \tag{3.4}$$

In formulas 3.1, 3.2 and 3.3, *tp* (true positive) and *tn* (true negative) are the number of the messages that have been correctly labelled as belonging to the positive and negative classes, respectively. On the other hand, *fp* (false positive) and *fn* (false negative) are the number of messages that have not been labelled as belonging to the negative and positive classes, respectively, but should have been. Accuracy is the proportion of correct labelled messages in the entire sample. Precision is the proportion of positive classifications that are actually correct and recall is the proportion of actual positives that are classified correctly. F1 is the harmonic mean of precision and recall [46].

The closer the F1-score is to 1, the better the performance of the model. If the performance is not acceptable, the sample size should be increased or a new method by using *bigrams* should be developed, for instance.

Chapter 4

Results and discussion

The main goal of this chapter is to present the results and analysis taking into consideration the three different collected datasets (London, Melbourne and New York). This work is based on real events from tweet messages so this analysis is important to get knowledge of its characteristics.

Next sections present the results of the model starting with the pre-processed data characterization where the messages are cleaned and rebuilt in order to be interpreted (4.1). Then, the two major sections are Text classification (4.2) and Sentiment Analysis (4.3).

Text classification analysis consists on understanding the performance results for each city considering the three sets of target words (small list (Table 3.2), medium list (Table 3.3), and big list (Table 3.4)). Then, messages are analysed according to temporal evolution and spatial distribution.

Sentiment analysis consists on collecting the emotions expressed on each message and analyse its evolution according to time and space. A specific analysis was made regarding a specific event on New York city.

4.1 Pre-processing

Table 4.1 presents an overview of the number of tweets collected, the number of tweets after the pre-processing phase, and a characterization of the metadata. Similar results were obtained among the three analysed cities. On average, 0.9% of the data was discarded after the pre-processing phase.

Table 4.1: Data characterization.

		New York	London	Melbourne
Total number of tweets		9 151 579	5 815 757	835 737
Total number of tweets after pre-processing		9 077 494 (99.2%)	5 762 023 (99.1%)	828 360 (99.1%)
Tweets with Hashtags (#)	Total (%)	2 361 811 (25.8%)	1 631 577 (28.1%)	257 755 (30.8%)
Tweets with User Mentions (@)	Total (%)	2 845 121 (31.1%)	2 652 870 (45.6%)	369 077 (44.2%)
Tweets with URLs	Total (%)	4 775 704 (52.2%)	3 054 889 (52.5%)	418 988 (50.1%)

About 30% of the tweets has hashtags and about 50% contain URL. The number of tweets with these special characters are similar among the three cities analysed. However, the number of tweets with user mentions are higher in London (45,6%) and Melbourne (44,2%), than in New York (31,1%), which denotes a different culture and environment.

4.2 Text classification

This section presents the performance of text classification model (section 4.2.1). Then, a temporal analysis of travel-related terms is first assessed (section 4.2.2), and then a spatial analysis is performed (section 4.2.3).

4.2.1 Performance

Table 4.2 presents the results of the performance of the BERT embeddings model for text classification using an *unigram* approach. As shown in Table 4.2, we obtained in general a high performance of results for the three assessed cities with the medium list 3.3. The average value of precision is 0.80, while accuracy and F1-score are both 0.90. The average value of Recall is the best achieved with an average value of 0.99. The results are consistent for the three cities assessed.

Table 4.2: Performance results of text classification model using three lists of travel-related terms for New York, London and Melbourne.

Performance measures	**New York**	**London**	**Melbourne**	**Average**
Small list of travel related words				
Precision	0.692	0.666	0.640	0.666
Recall	0.906	0.920	0.882	0.902
Accuracy	0.810	0.804	0.777	0.797
F1-score	0.785	0.773	0.742	0.766
Medium list of travel related words				
Precision	0.824	0.816	0.788	0.809
Recall	0.990	0.988	0.992	0.990
Accuracy	0.908	0.903	0.891	0.901
F1-score	0.900	0.894	0.878	0.891
Big list of travel related words				
Precision	0.634	0.578	0.604	0.605
Recall	0.984	0.986	0.993	0.988
Accuracy	0.812	0.785	0.800	0.799
F1-score	0.771	0.729	0.751	0.751

For the other lists 3.2 and 3.4, the performance results were similar and, in general, worse than the medium list. This happened because the small list is too short and incomplete, so BERT model is not able to label a great amount messages as travel related and the number of *fn* (false negatives)

messages is high. The big list has a big amount of ambiguous words that make the number of *fp* (false positives) messages increase comparing to the medium list.

Text classification systems for social media face a lot of challenges. The main challenge is certainly dealing with large vocabularies once the lexical variation caused by informal text is significant and the messages have small sizes. This aspect requires some dexterity to manage and control the problem by getting a solution in the pre-processing stage outlining the lexical variation.

The performance evaluation of the model was good, despite some false positive and negative labels. Due to using an *unigram* approach, some words listed on Table 3.3 were wrongly associated with mobility, such as *train*, *station* and *underground*. Beside the ambiguity of some terms also used in other domains, the performance was generally good.

Some words such as *street*, *walk*, *train* and *car* are the most common words in travel-related tweets in the three cities. There are some other words with non-significant number of references, such as *motorcycle* and *highway*. Therefore, this suggests that for further studies, namely real-time analysis, these words could be discarded.

A benchmark analysis shows that the results obtained are in agreement with the ones obtained in the literature 2. With the BERT embeddings model better results were obtained than the ones yielded with supervised machine learning techniques for classification, such as support vector machines, logistic regression, and random forest, namely using BoW [12].

4.2.2 Temporal analysis

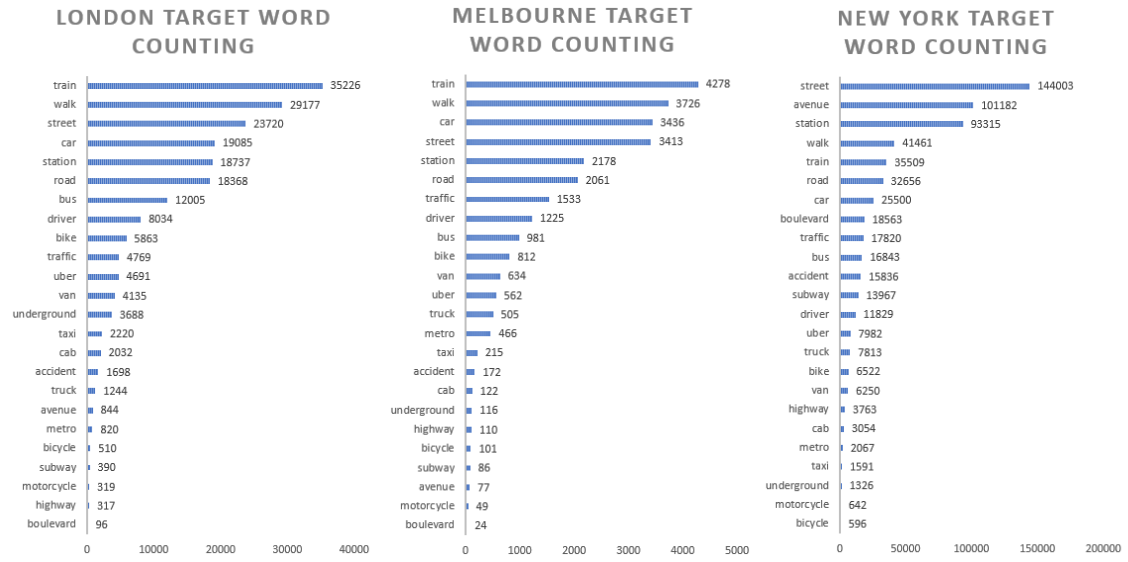
Table 4.3 shows an overview of a set of indicators for each analysed city. Among the three analysed cities, New York is the one that has the highest average number of daily messages (approximately 93 000 daily tweets). The average period with the most messages exchanged is in the evening (20-21h); however, no pattern related to the day of the week was identified.

Table 4.3: Indicators.

Characteristics	**New York**	**London**	**Melbourne**
File size (Gb)	28.4	18.60	2.6
Average number of daily tweets	92 304	60 187	8 237
Popular days of the week	Thursday	Wednesday	Sunday
Popular hour of the day	21	21	20
Percentage of travel related tweets	4.82%	3.21%	3.20%

Regarding traffic-related messages collected with the medium list, we found 440.877 tweets (5%) in New York, 186.251 tweets (3%) in London, and 26.744 tweets (3%) in Melbourne matching the terms selected in Table 3.3. While in London and Melbourne the most popular words analysed were *train*, *walk*, and *street*, in New York, the most popular words were *street*, *avenue* and *station*. Figure 4.1 shows the number of messages with travel-related terms for the three cities.

The number of travel related messages is in accordance with other studies (e.g., [12]).



(a) (b) (c)

Figure 4.1: Number of messages with travel-related terms from the medium list 3.3 for London (a), Melbourne (b) and New York (c).

The analysis of the distribution of travel-related messages along the day shows similar patterns among the analysed cities. The percentage of traffic-related messages sent during the typical working hours of the day is around 5-6%. Over this period of the day, the percentage of messages does not change significantly. The maximum number of messages is reached at 4PM in New York (6.0%), and at 5PM in London (6.5%) and at 6PM in Melbourne (7.6%). According to the TomTom traffic congestion index, these values are coincident with the afternoon peak hour or are one hour after the peak hour [5] (Figure A.3). Figure 4.2 shows the percentage of travel-related messages along the day for the three cities analysed.

Figure 4.3 and Figure 4.4 show the average number of tweets per hour of the day and day of the week respectively, for each list of travel-related terms (small, medium and big) and city (New York, London and Melbourne).

It is possible to recognize a pattern in the Figures 4.2 and 4.3 where the common working hours are the ones when the flow of twitter messages is higher. For each city and each list of target words, the pattern is consistent. For the bigger list, the pattern is slightly better marked. Figure 4.3 shows a hourly deviation of around 35% of the mean value. The maximum standard deviation is observed during the afternoon peak-hour.

Regarding the days of the week, there were no significant patterns recognized, so it may not be a great influence on travel related messages on Twitter. However, it shows that on weekend days the number of travel related tweets is slightly lower than on the other week days. Figure 4.4 shows the total number of travel related tweets per day of the week.

The patterns recognized in this analysis are consistent with the patterns identified in TomTom index (Figure A.3). The peak hours are from 5 a.m. to 9 a.m. and from 3 p.m. to 8 p.m. in the

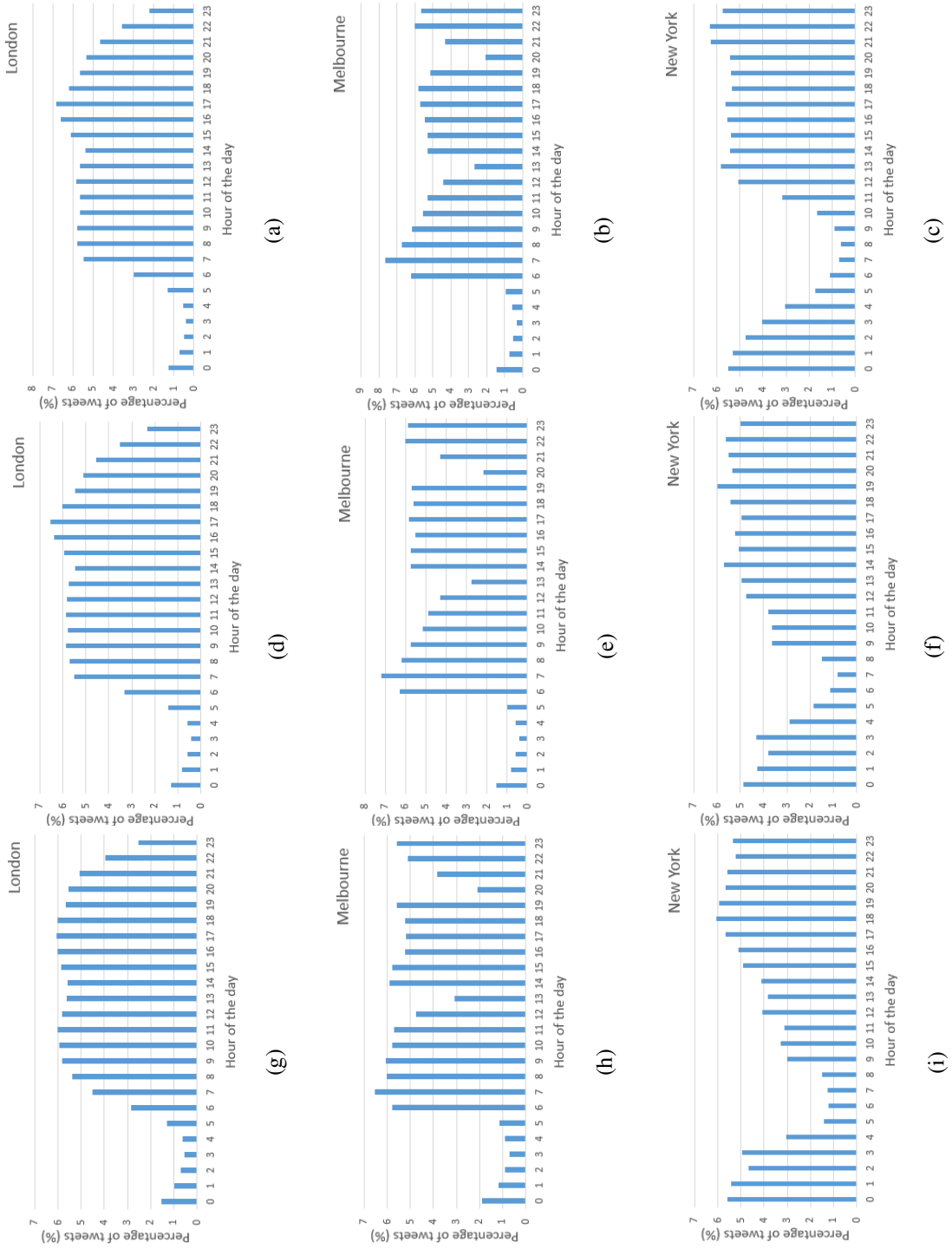


Figure 4.2: Percentage of tweets per hour of the day of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i).

three cities. Figure A.4 shows the evolution of hourly average traffic congestion on 2019, 2020 and along the week of September 14, 2021, and it shows that the patterns remain the same, even in more recent years

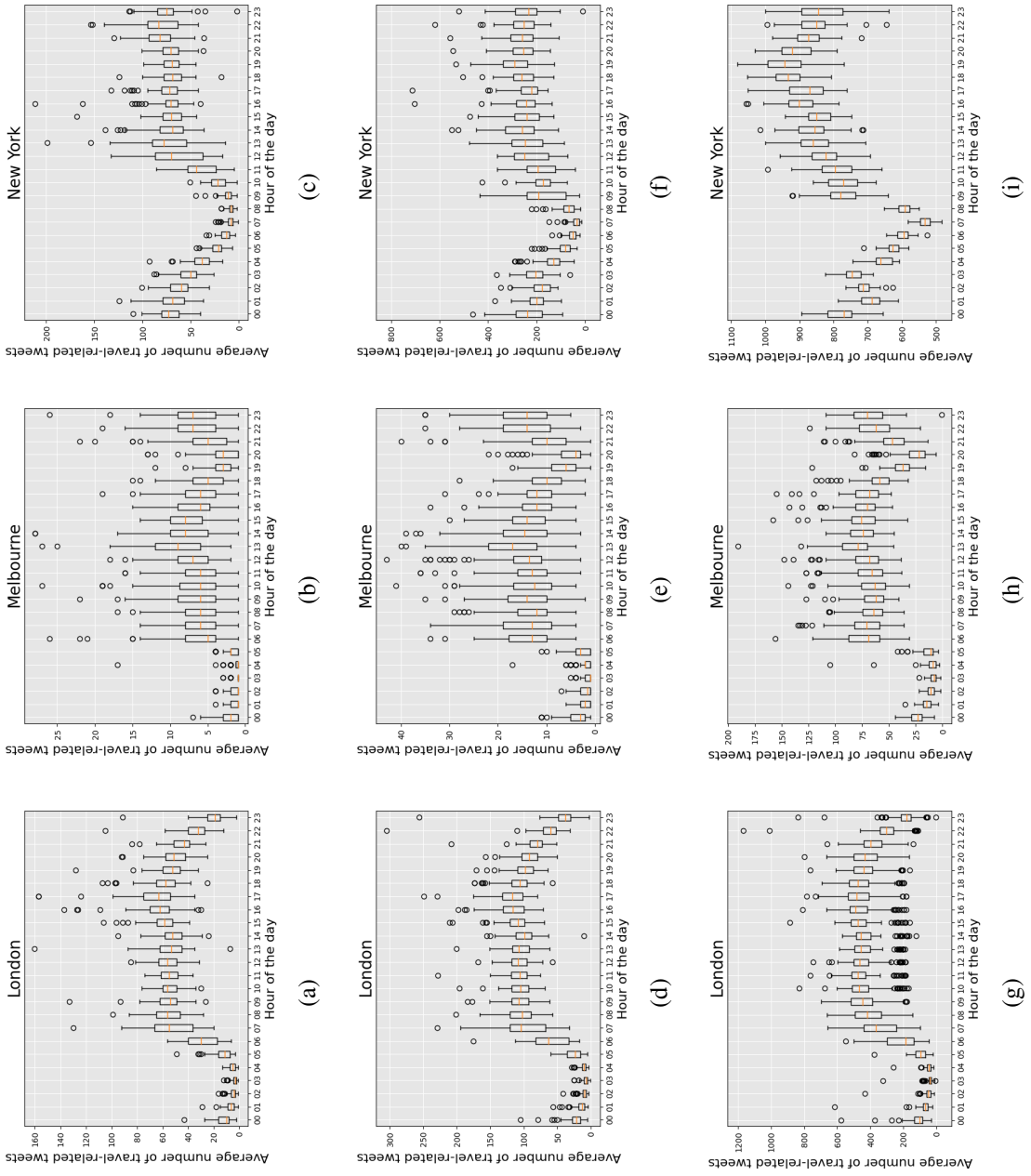


Figure 4.3: Average number of tweets per hour of the day of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i).

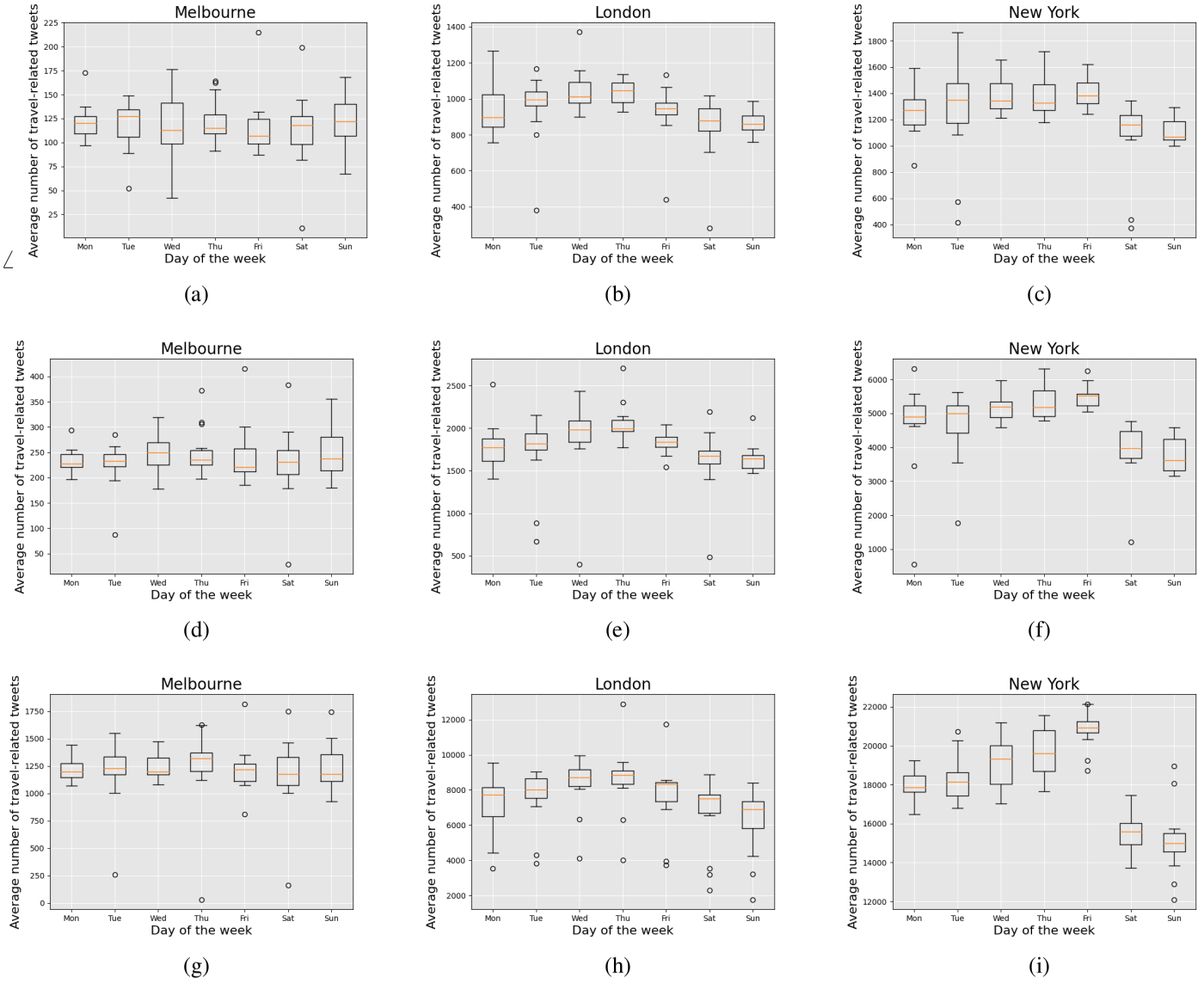


Figure 4.4: Average number of tweets per day of the week of London (a, d, g), Melbourne (b, e, h) and New York (c, f, i) related to the small list 3.2 (a, b, c), medium list 3.3 (d, e, f) and big list 3.4 (g, h, i).

4.2.3 Spatial analysis

In order to apply a spatial analysis to the data, coordinates should be analysed. Some tweets, instead of having specific coordinates assigned, are associated to bounding-boxes. Then, an approximated coordinate was assigned to those tweets by calculating the middle point of the bounding-boxes.

Figure 4.5 represents the geographical distributions of total and travel related tweets from London, Melbourne and New York, according to the medium list of target words 3.3, which is the list with best performance results. There is a sharp difference between the total tweets plots and travel related tweets plots. The travel related plots show less density than total tweets plots because of the BERT model filtering performed. Although the performance results where high, the difference would be sharper if the performance was perfect.

The geographical plot is important to detect the most affluent zones in each city. It is clear that the centres of each city are the zones with the most number of tweets and it corresponds to the zones with more population density. For example, in the city centre of London, the closer it gets to the river, the more dense the plot gets, once it is one of the most populated zones in London. According to [60], around 9 million people go by London city center daily.

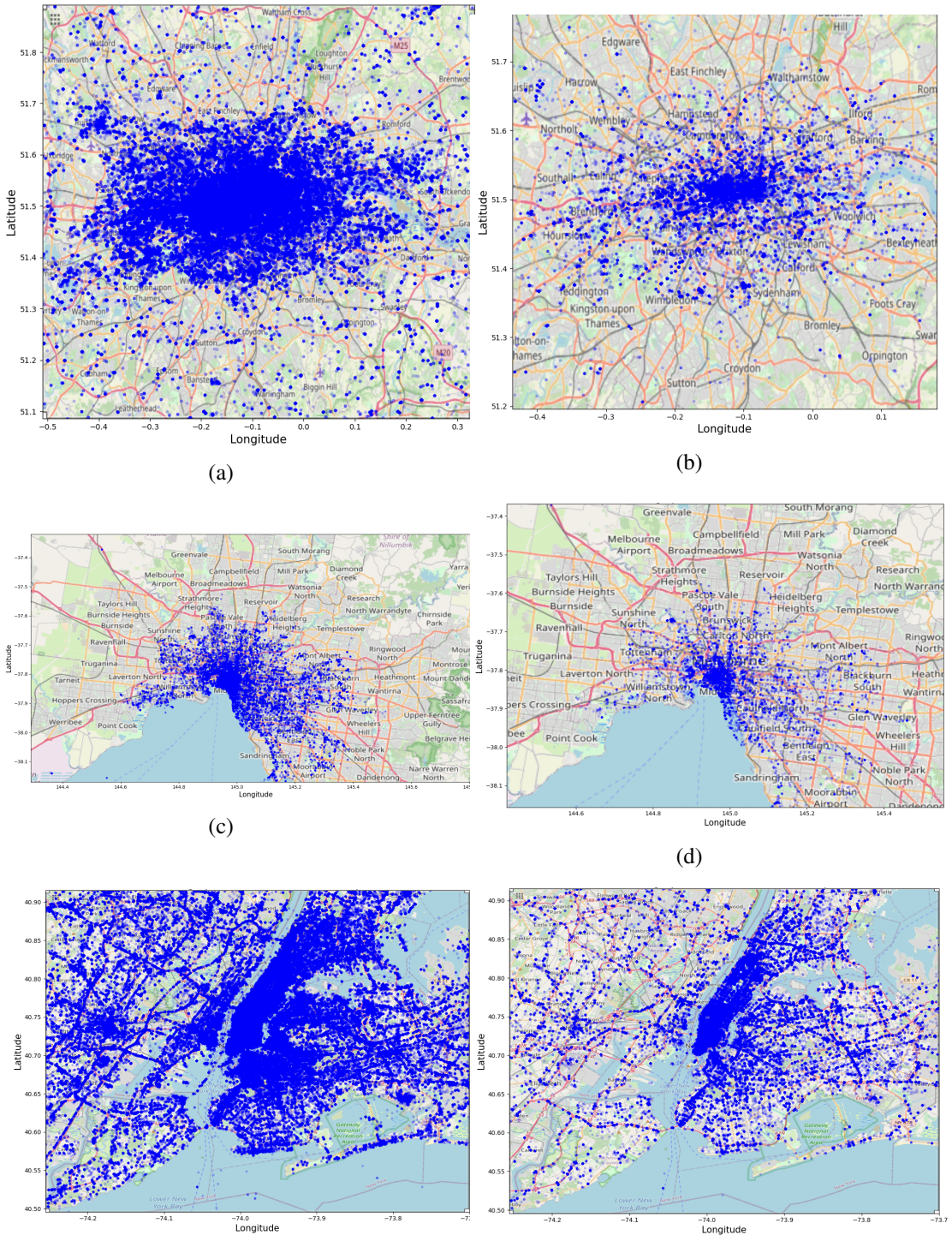


Figure 4.5: Geographical distributions of tweets for London (a), Melbourne (c) and New York (e) and geographical distributions of travel related tweets for London (b), Melbourne (d) and New York (f) according to the medium list of target words 3.3.

4.3 Sentiment analysis

This section presents the performance of text classification model (section 4.3.1). Then, a temporal analysis of travel-related terms is first assessed (section 4.3.2), and then a spatial analysis is performed (section 4.3.3).

4.3.1 Performance

Table 4.4 shows the performance results of VADER sentiment model for the medium list of target words (Table 3.3) which obtained the best performance results for the BERT model. The performance shows good results for the three assessed cities where London got the best ratings with recall, accuracy and F1-score reaching values over 0.80. The average value of precision scored 0.77 while recall, accuracy and F1-score attained around 0.78. The results demonstrate consistency along the three assessed cities.

Table 4.4: Performance results of sentiment analysis using the medium list of travel-related terms 3.3 for London, Melbourne and New York.

Performance measures	London	Melbourne	New York	Average
Precision	0.720	0.790	0.800	0.770
Recall	0.758	0.849	0.741	0.780
Accuracy	0.745	0.825	0.760	0.777
F1-score	0.738	0.819	0.769	0.775

4.3.2 Temporal analysis

Figure 4.6 and Figure show the average number of traveled-related tweets per hour of the day and day of the week respectively. The data are considering the sentiment (positive, negative and neutral) for London, Melbourne, and New York, according to the medium list of target words 3.3.

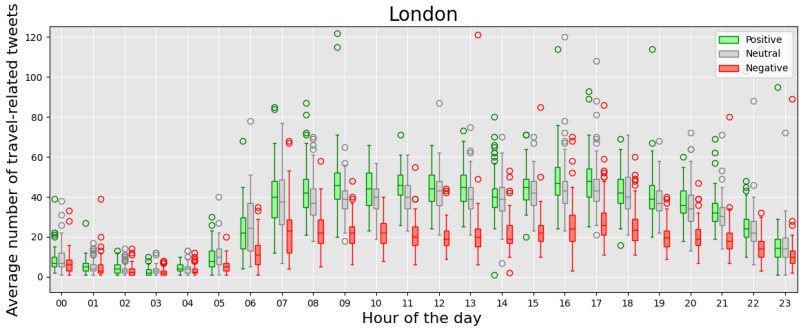
As can be observed, the number of positive and neutral traveled-related tweets are almost equal in London and Melbourne. However, in New York the number of neutral traveled-related tweets is higher than the number of positive traveled-related tweets. In the three cities analysed the number of negative traveled-related tweets is around 50% of the number of positive traveled-related tweets.

Regarding the hourly variation of travel-related tweets, Figure 4.6 shows the average number of traveled-related tweets per hour of the day considering the sentiment (positive, negative and neutral) for London, Melbourne, and New York. The variations are consistent for each hour of the day and represent around 80% of the mean value. The patterns recognized on Figure 4.3 are also confirmed in this analysis.

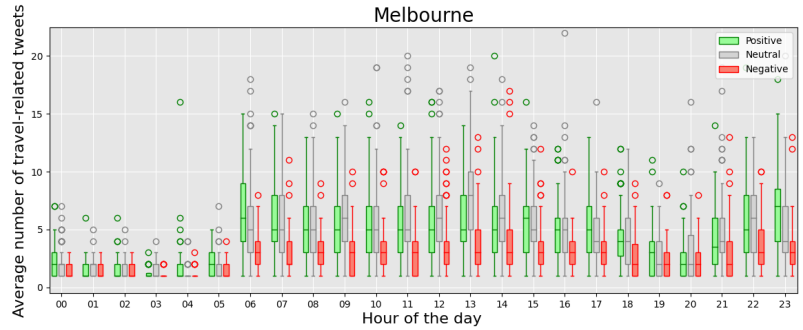
Along the week, Wednesday and Thursday are the most popular weekdays. The number of positive and negative travel-related tweets hits, during these days, the maximum values of the week in all the three cities analysed (Figure 4.7). This might suggest that these days are crowded

and have a lot of traffic. In fact, figure A.3 shows that Thursday is one of the most congested day of the week, according to TomTom index.

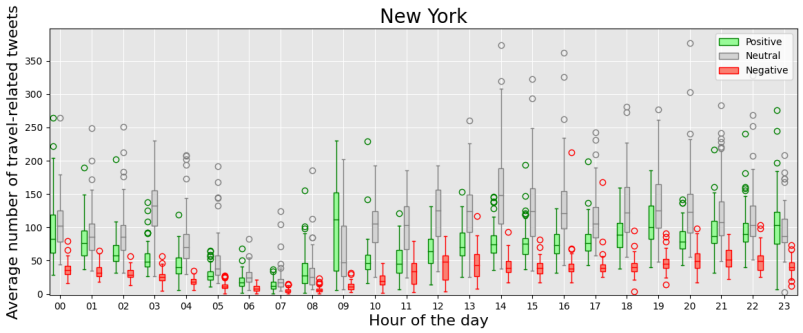
Weekends are in general the periods of the week with less number of traveled-related tweets. This trends is clear observed in the biggest cities of London and New York, but not in Melbourne, which can be related with the low number of traveled-related tweets recorded daily in this city, around 100 tweets positives and around 50 tweets negative. With exception of London, the standard deviation of the positive travel-related tweets is higher during weekend. However, this pattern is not observed for the negative travel-related tweets.



(a)



(b)



(c)

Figure 4.6: Average number of tweets per hour of the day considering the sentiment for London (a), Melbourne (b) and New York (c), according to the medium list of target words 3.3.

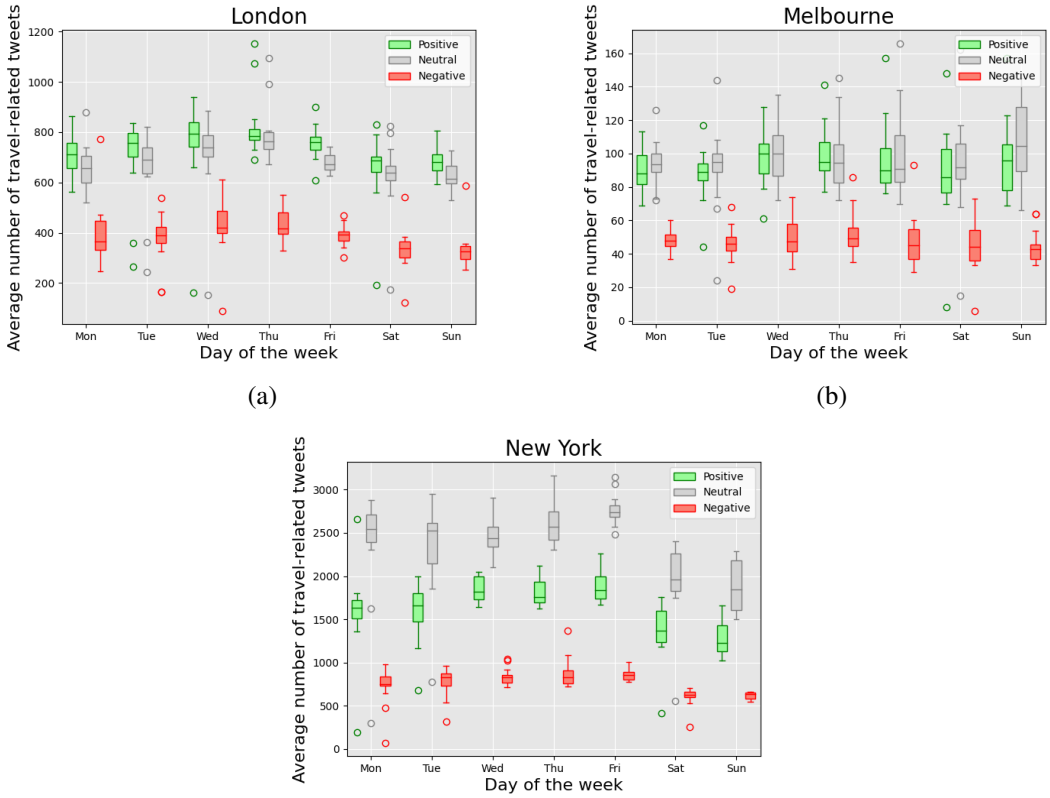


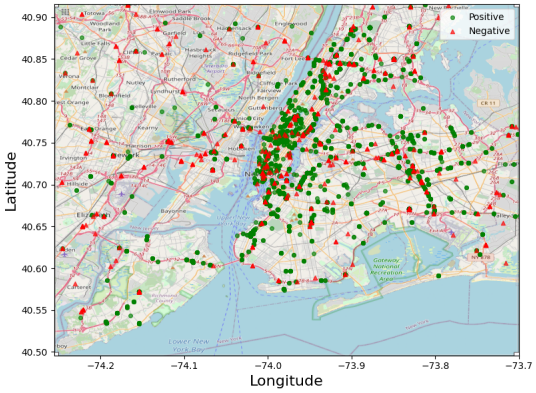
Figure 4.7: Average number of tweets per day of the week considering the sentiment for London (a), Melbourne (b) and New York (c), according to the medium list of target words 3.3.

4.3.3 Spatial analysis

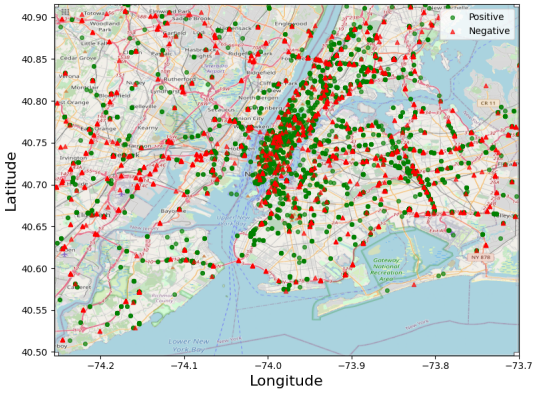
To analyse the distributions of travel-related tweets, maps of positive and negative tweets were produced for the city of New York, for May 17 (Thursday) from 2017 (plots with the neutral messages are on the Figures A.1 and A.2). Figure 4.8 plots that distribution from 6 a.m. to 12 a.m.. These maps were computed for the results of the model obtained using the medium list of target words (Table 3.3), as this list was the one with the best model performance.

As can be observed, the distribution of travel-related terms change along the day. The density of messages grows along the day. It is possible to see that in the morning (6 a.m. to 12 p.m.), there are fewer messages identified than for the rest of the day. As it gets closer to 9 p.m., the density of tweets grows considerably until it gets lighter from 9 p.m. to 12 a.m..

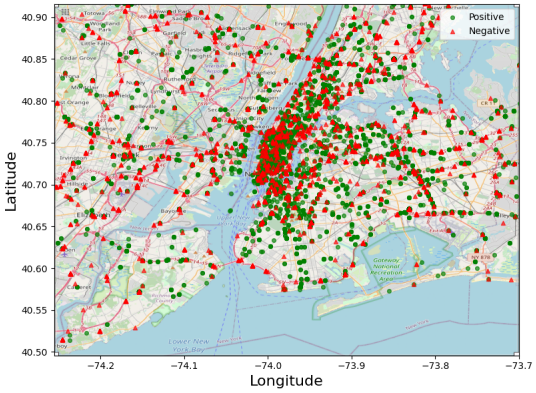
Regarding the sentiment, as the density of the messages grows, the more negative messages appear overtaking the number of positives. From 6 a.m. until 12 p.m., the number of positive messages is greater or equal to the number of negatives. Once the affluence to the streets gets higher, the number of negative messages is bigger than the number of positives, as it is seen in Figure 4.8 from 12 p.m. to 9 p.m..



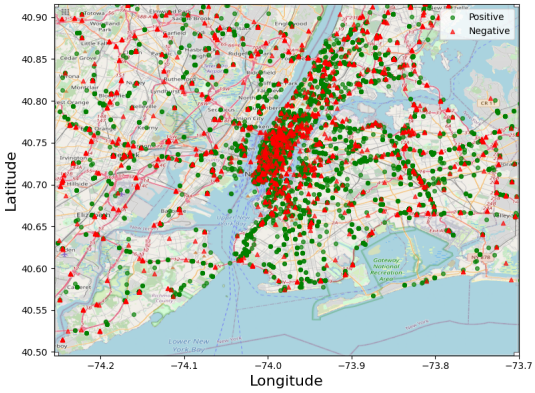
(a) 6 a.m. to 9 a.m.



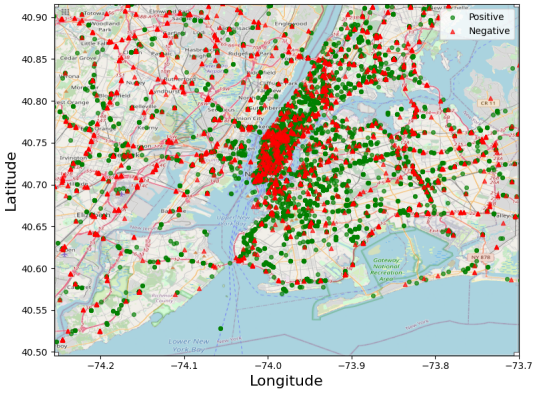
(b) 9 a.m. to 12 p.m.



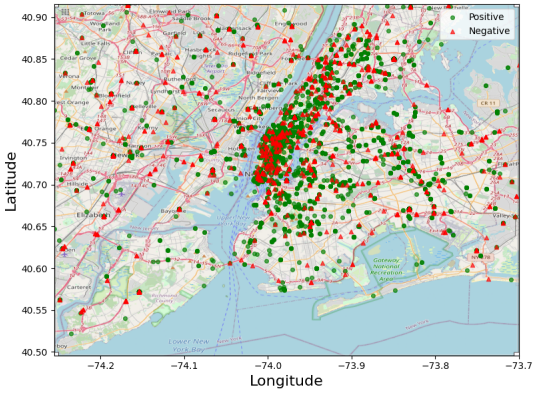
(c) 12 p.m. to 3 p.m.



(d) 3 p.m. to 6 p.m.



(e) 6 p.m. to 9 p.m.



(f) 9 p.m. to 12 a.m.

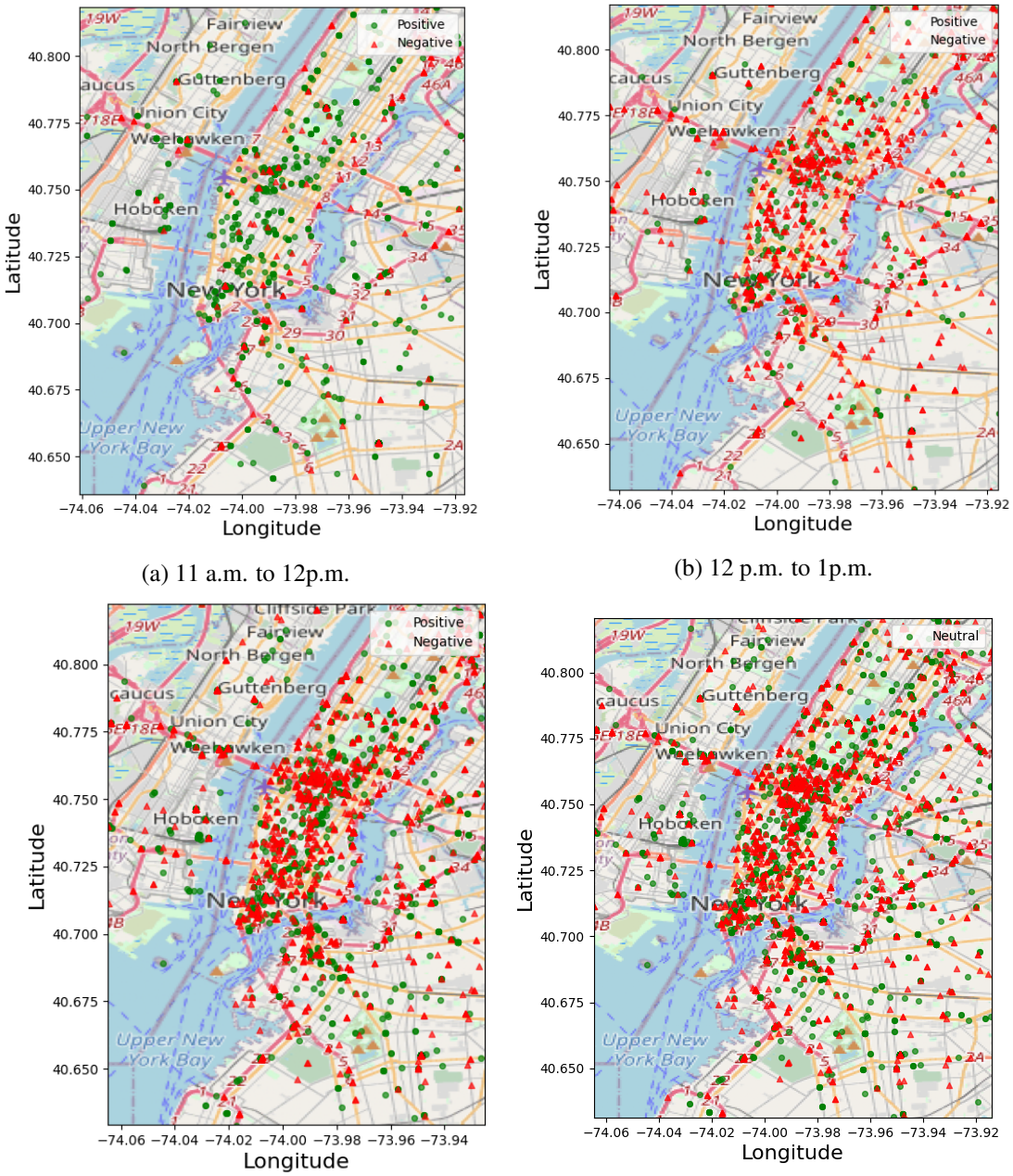
Figure 4.8: Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment from 6a.m. to 12a.m., according to the medium list of target words (Table 3.3).

In order to go further in the spatial analysis of the model, a day with a crash in the city of New York was analysed. According with [2] at 11:53 EDT of May 18, 2017, a car was crashed in Times Square, New York City, United States. One person was killed and 20 were injured. To analyse this occurrence, the model was plotted over Manhattan in the hour before and two hours after of the accident occur. Figure 4.9 shows the location of the accident, while Figure 4.10 presents the spatial distribution of positive and negative travel-related terms over Manhattan (New York) in the previous hour of the accident and the next two hour.



Figure 4.9: 2017 Times Square car crash (Manhattan, New York) [2].

During the hour before the crash, the flow of messages was low and mostly with positive sentiment, as it is on Figure 4.10a. Right after the hour of the crash, the density of messages grows with a low number of positive messages and a clear advantage for the negative. On Figures 4.10b, 4.10c and 4.10d, a red stain is all over Manhattan which can be related to the crash related on [2]. Moreover, three hours after the accident the number of negative messages increases even more, which suggest that a general increase of traffic congestion happens in the city caused by the occurred accident.



(a) 11 a.m. to 12p.m.

(b) 12 p.m. to 1p.m.

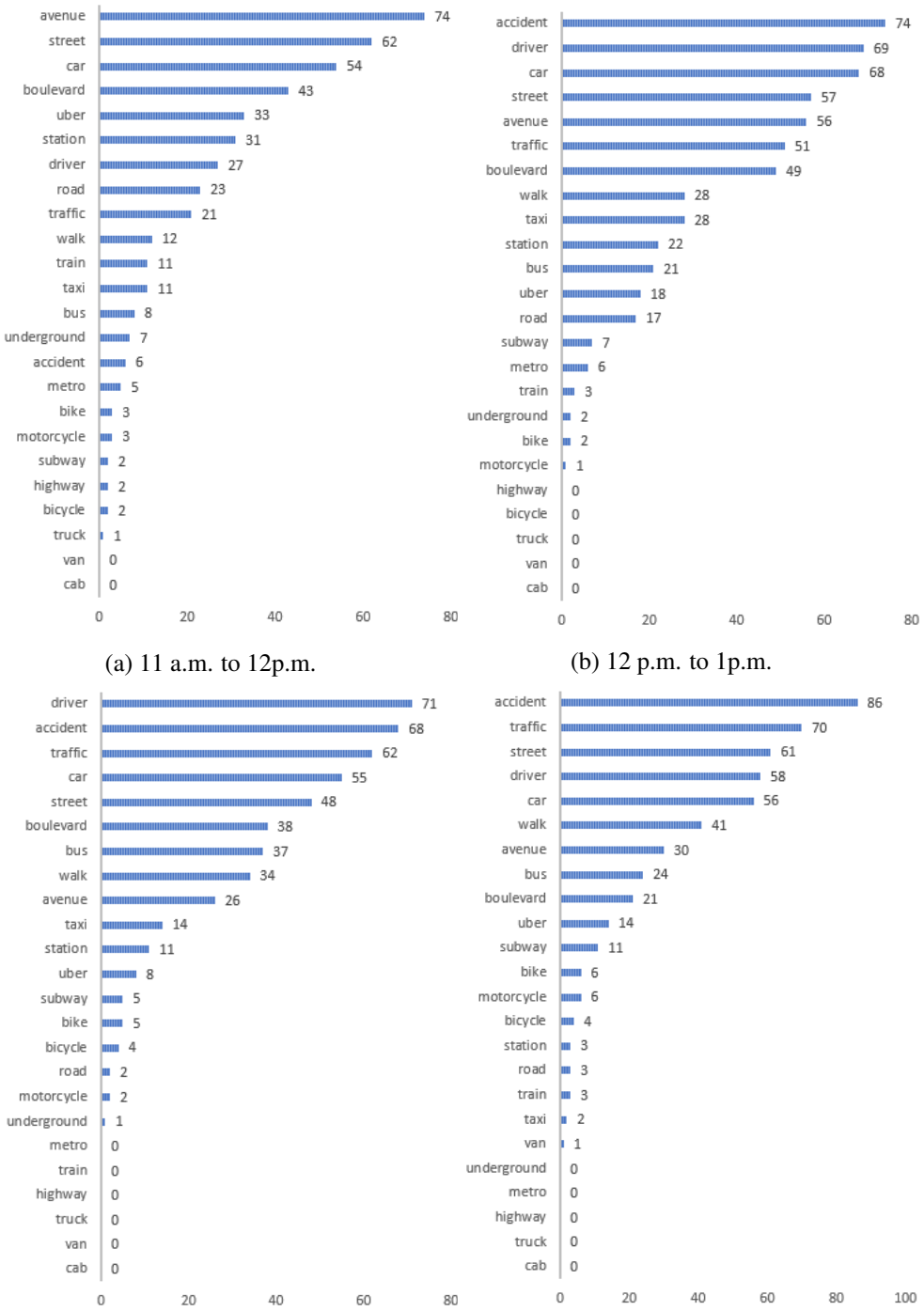
(c) 1 p.m. to 2 p.m.

(d) 2 p.m. to 3 p.m.

Figure 4.10: Density and sentiment evolution of the messages one hour before and three hours after a car was crashed in Times Square, New York City, United States, at 11:53 EDT of May 18, 2017.

Figure 4.11 represents the number of travel related terms from the medium list 3.3 in New York for one hour before the accident and three hours after. The word *accident* needs to be highlighted because from 11 a.m. to 12 p.m. it was used 6 times, but after the crash, from 12 p.m. to 13 p.m., it was the most used word with 74 appearances. Some terms related to the accident are also present in the group of the most used words after the crash, such as, *driver*, *car* and *street*. Relating to the sentiment, one word that appeared mostly after the crash and has a negative impact (beside the

word *accident*) is the word *traffic*.



(a) 11 a.m. to 12p.m.

(b) 12 p.m. to 1p.m.

(c) 1 p.m. to 2 p.m.

(d) 2 p.m. to 3 p.m.

Figure 4.11: Number of messages with travel-related terms from the medium list 3.3 for New York one hour before the accident and three hours after.

Chapter 5

Conclusions and future work

In this chapter, the main conclusions are mentioned (section 5.1) as well as the publications (sections 5.2) and the suggestions for future work (section 5.3).

5.1 Conclusions

In this work, we extracted knowledge from social media to evaluate urban transport. For this purpose, we used BERT embeddings with an *unigram* approach and VADER sentiment to assess the transportation sector in three major worldwide cities: New York, London and Melbourne. For this purpose, a methodology structured into three main modules was defined: system configuration, data analytics, and visualization.

To demonstrate the methodology proposed, first, the data extraction and its pre-processing were developed with NLP techniques (using NLTK tool). Pre-processing is a very important stage of this work once texts extracted from Twitter are full of informal words, slang and misspellings.

Then, the classification with BERT embeddings was implemented and analysed. For that, three lists of target words were defined small list 3.2, medium list 3.3 and big list 3.4). A higher performance of the model was consistently obtained for the three cities studied. The performance results were good and consistent with the benchmark analysis in the literature 2. The medium list of travel related target words showed the best performance results because is the less ambiguous and the more balanced out of the three lists. The results show that the methodology used is able to identify trends and therefore can be potentially used to support the analysis of traffic-related events.

Sentiment analysis is considered a good way to assess about public satisfaction regarding transportation and mobility. It is important to recognize public needs and to satisfy them. In this work, the sentiment analysis was performed by using a lexicon and rule-based tool called VADER. VADER sentiment attained good results for the three assessed cities according to the medium list of travel related words. Some satisfaction analysis was made for the days of the week and for the hours of the day. The results show that, beside there is a greater number of positive messages, the average number of negative messages is significant. A random working day of the week was

chosen to measure public satisfaction on transportation in New York. It is possible to infer that as the affluence to the streets grow, the dissatisfaction grows along. A similar analysis was made for a specific event on May 18, 2017 in New York when a car crash happened in Manhattan around mid day. During the hour before the crash, traffic satisfaction was moderate, but as soon as the crash happened, the dissatisfaction increased drastically during the three hours after.

With the general growth of cities, the need for mobility increases. As a result, the levels of traffic congestion have been raising dramatically worldwide. Passenger transit is projected to increase, with respect to 2005, by around 34% in 2030 and 51% by 2050 [5]. In Europe, congestion costs are about 1% of gross domestic product (GDP) each year [6]. According to the White Paper on transport (2011), the congestion costs will increase by about 50% by 2050. As a consequence, the social costs of accidents, noise, and pollution will continue to raise.

According to this, crowd sensing will be more and more important to reduce costs, time, and also to complement traditional surveying methods. Although both present some evident differences, they also offer compatible and complementary characteristics. Although Twitter users are younger, more affluent, and more educated than the general population as a whole, the methodology here presented is a good complement to urban planning measures and to traffic control statistics, at least for large cities with large populations where the amount of data is big enough to get results. Therefore, the work developed demonstrate that the methodology defined can be used in a tool to assist transport engineers, urban planners, researchers, and policymakers in getting insight into public opinions regarding urban mobility.

5.2 Publications

During the developing of this dissertation, one scientific paper was presented in a workshop at the *IEEE International Smart Cities Conference 2021* in order to share the methodology and its results:

- Francisco Murços, Tânia Fontes, Rosaldo Rossetti. Are BERT embeddings able to infer travel patterns from Twitter efficiently using a unigram approach? In *IEEE International Smart Cities Conference* (IEEE ISC2), 2021. To appear.

Another scientific paper about this dissertation is being prepared with the most recent results and the addition of sentiment analysis.

5.3 Future work

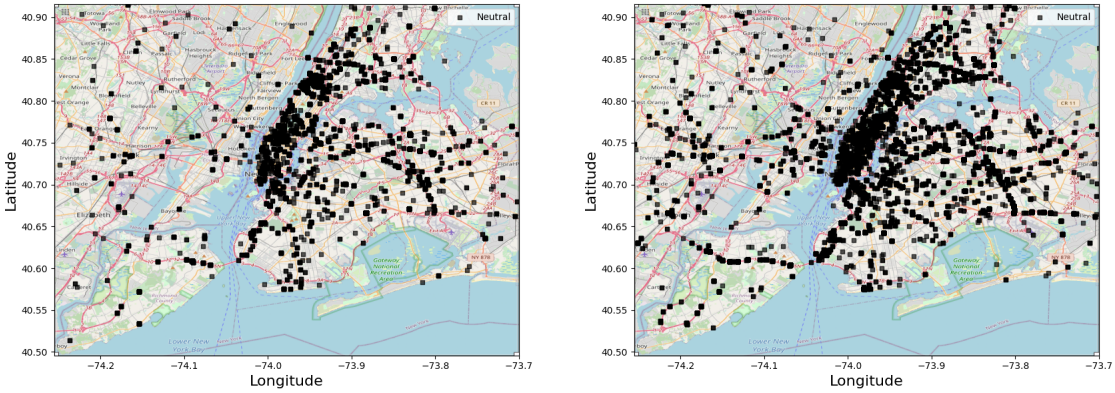
The dissertation proposed an innovative methodology that leads to further research work. The methodology proposed has great potential and proves how people are increasingly connecting more and more to each other and to every kind sector, in this case, to transportation and mobility

sector. Taking into considerations some suggestions might develop the area of artificial intelligence in transportation and mobility and improve the knowledge extracted from social media. Here are some suggestions that come in sequence of this work:

- Using pre-trained BERT large model (uncased) could lead to better and more reliable performance results. The BERT large model is an embedding model with twice the embedding layers compared to BERT base model. It is an extensive time consuming task but could lead to new information;
- Fine-tuning a BERT model would be another option to explore BERT embeddings and get more results regarding transportation and mobility;
- In this work, an *unigram* approach was used which means that each token for the embedding was a single word. Another option would be to rebuild the proposed lists of target words and use an *n-gram* approach higher than *unigram*. It would capture even better the context of the expressions;
- It would be interesting to apply this methodology to a language different than *English*. There are more large cities with large populations around the world where the spoken language is not *English*. Figueiredo [12] applied his methodology to two Brazilian cities, Rio de Janeiro and São Paulo, and it would be interesting to apply BERT embeddings to these cities and do some comparison research;
- Regarding the sentiment analysis, fine-tuning VADER sentiment model and improving its lexicon could lead to more reliable sentiment results;
- An interesting study would be to analyse the sensibility of VADER, TextBlob and Affin sentiment analysis models by comparing its performances' results in the mobility and transportation sector;
- Nowadays, information is at the distance of a click. Considering it, an approach that could be able to analyse Twitter texts' emotion on transportation and mobility on real-time would be a welcome addition to the methodology.

Appendix A

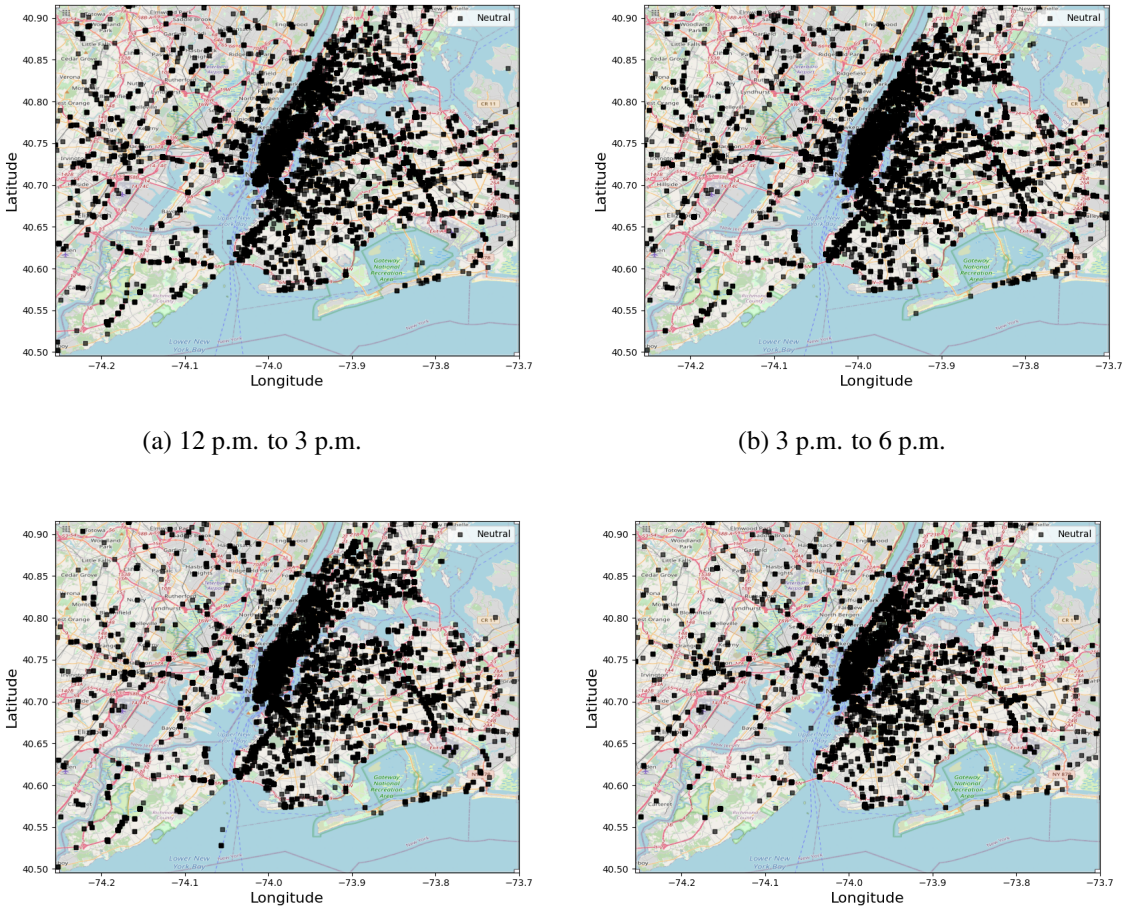
A.1 Plots with the neutral messages for New York on May 17, 2017



(a) 6 a.m. to 9 a.m.

(b) 9 a.m. to 12 p.m.

Figure A.1: Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment (neutral) from 6 a.m. to 12 p.m., according to the medium list of target words (Table 3.3).



(a) 12 p.m. to 3 p.m.

(b) 3 p.m. to 6 p.m.

(c) 6 p.m. to 9 p.m.

(d) 9 p.m. to 12 a.m.

Figure A.2: Messages distributions for New York on May 17 from 2017 (thursday) considering the sentiment (neutral) from 12 p.m. to 12 a.m., according to the medium list of target words (Table 3.3).

A.2 Statistical data regarding transportation on London, Melbourne and New York from TomTom index

	Sun	Mon	Tue	Wed	Thu	Fri	Sat
12:00 AM	19%	6%	5%	7%	9%	12%	18%
	13%	2%	1%	2%	3%	6%	12%
02:00 AM	10%	0%	0%	0%	1%	3%	8%
	7%	0%	0%	0%	0%	0%	5%
04:00 AM	3%	0%	0%	0%	0%	0%	2%
	0%	1%	1%	1%	1%	1%	0%
06:00 AM	0%	23%	22%	22%	21%	17%	1%
	2%	52%	53%	52%	50%	39%	6%
08:00 AM	5%	63%	65%	64%	61%	49%	11%
	11%	43%	46%	46%	45%	39%	20%
10:00 AM	19%	34%	37%	38%	38%	37%	29%
	28%	34%	35%	38%	38%	40%	39%
12:00 PM	37%	35%	36%	39%	40%	44%	46%
	39%	34%	36%	39%	39%	45%	47%
02:00 PM	37%	35%	36%	40%	41%	49%	42%
	34%	43%	46%	50%	52%	60%	38%
04:00 PM	35%	49%	54%	59%	62%	66%	37%
	37%	55%	63%	68%	71%	67%	39%
06:00 PM	36%	45%	54%	59%	61%	56%	38%
	29%	28%	34%	37%	40%	40%	33%
08:00 PM	21%	18%	21%	23%	25%	28%	27%
	16%	14%	18%	19%	20%	22%	22%
10:00 PM	13%	14%	18%	20%	22%	23%	23%
	10%	10%	14%	16%	19%	22%	22%

(a) London

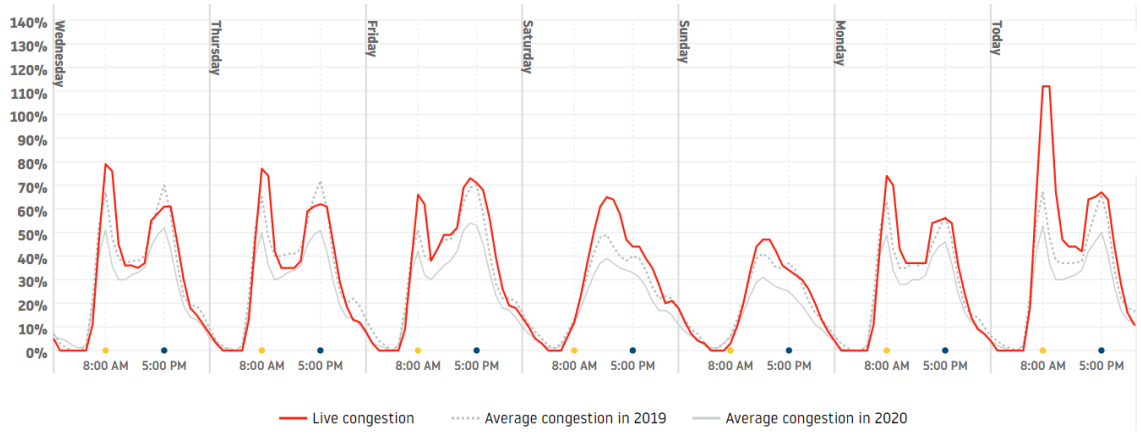
	Sun	Mon	Tue	Wed	Thu	Fri	Sat
12:00 AM	13%	3%	3%	3%	4%	5%	10%
	9%	0%	0%	0%	1%	2%	7%
02:00 AM	6%	0%	0%	0%	0%	0%	4%
	4%	0%	0%	0%	0%	0%	2%
04:00 AM	1%	0%	0%	0%	0%	0%	0%
	0%	2%	3%	3%	3%	2%	0%
06:00 AM	1%	18%	22%	22%	22%	18%	2%
	2%	37%	42%	42%	41%	32%	5%
08:00 AM	5%	54%	59%	59%	57%	45%	11%
	10%	32%	37%	38%	38%	31%	17%
10:00 AM	15%	23%	27%	28%	28%	27%	23%
	20%	23%	26%	27%	28%	29%	29%
12:00 PM	25%	24%	27%	29%	30%	31%	34%
	23%	23%	26%	27%	29%	30%	31%
02:00 PM	21%	25%	28%	30%	31%	35%	27%
	20%	37%	43%	46%	48%	53%	24%
04:00 PM	21%	38%	47%	51%	54%	57%	23%
	20%	45%	56%	60%	64%	57%	26%
06:00 PM	17%	30%	39%	43%	46%	39%	25%
	13%	14%	18%	20%	22%	23%	21%
08:00 PM	11%	12%	13%	14%	16%	16%	17%
	11%	11%	13%	14%	15%	16%	16%
10:00 PM	9%	8%	10%	12%	13%	16%	18%
	6%	6%	6%	8%	9%	14%	17%

(b) Melbourne

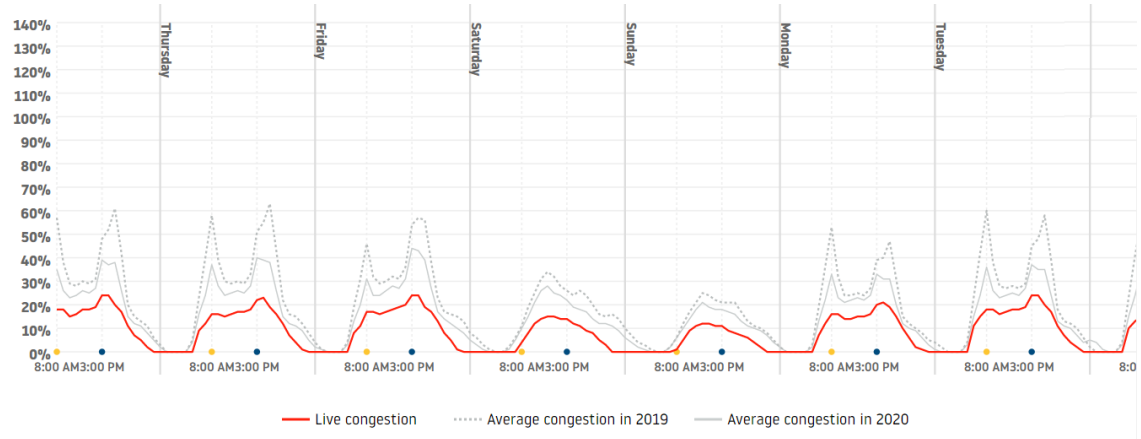
	Sun	Mon	Tue	Wed	Thu	Fri	Sat
12:00 AM	18%	7%	7%	10%	11%	14%	18%
	12%	4%	4%	5%	6%	8%	13%
02:00 AM	8%	2%	2%	3%	3%	4%	8%
	5%	0%	1%	1%	1%	2%	5%
04:00 AM	4%	0%	0%	0%	0%	1%	4%
	1%	5%	5%	5%	5%	5%	2%
06:00 AM	0%	26%	27%	28%	27%	23%	3%
	0%	48%	50%	50%	50%	41%	6%
08:00 AM	3%	54%	57%	56%	57%	46%	11%
	8%	39%	42%	41%	44%	36%	17%
10:00 AM	14%	31%	34%	34%	36%	32%	22%
	21%	30%	32%	33%	35%	34%	30%
12:00 PM	27%	29%	31%	32%	35%	36%	36%
	33%	29%	32%	33%	36%	40%	38%
02:00 PM	37%	35%	38%	41%	44%	53%	40%
	39%	44%	50%	54%	58%	69%	42%
04:00 PM	40%	53%	61%	62%	66%	76%	44%
	40%	58%	67%	69%	73%	77%	44%
06:00 PM	37%	46%	54%	57%	61%	62%	42%
	33%	27%	33%	35%	40%	42%	36%
08:00 PM	30%	19%	23%	23%	29%	30%	30%
	26%	15%	19%	19%	24%	24%	26%
10:00 PM	19%	12%	16%	17%	22%	22%	24%
	12%	10%	14%	15%	19%	21%	22%

(c) New York

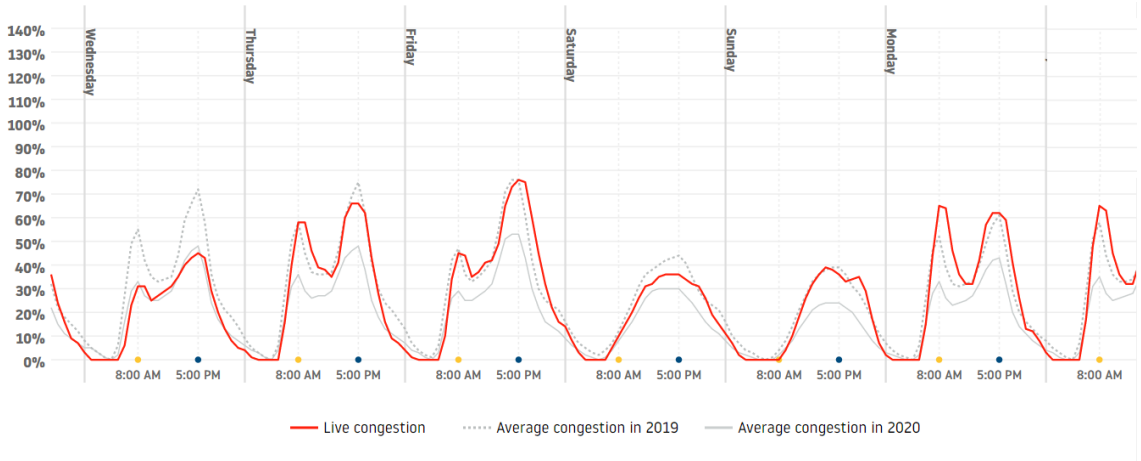
Figure A.3: Average traffic congestion in 2017 for London (a), Melbourne (b) and New York (c) according TomTom index.



(a) London



(b) Melbourne



(c) New York

Figure A.4: Hourly average traffic congestion along the week of 14/09/2021 for London (a), Melbourne (b) and New York (c) according TomTom index.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[2] Wikipedia contributors. 2017 times square car crash — Wikipedia, the free encyclopedia, 2021. [Online; accessed 15-September-2021]. URL: https://en.wikipedia.org/w/index.php?title=2017_Times_Square_car_crash&oldid=1034609088.

[3] Ayelet Gal-Tzur, Susan M Grant-Muller, Tsvi Kuflik, Einat Minkov, Silvio Nocera, and Itay Shoor. The potential of social media in delivering transport policy goals. *Transport Policy*, 32:115–123, 2014.

[4] UNDESA. *World Urbanization Prospects*, volume 12. 2018. URL: https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf.

[5] TOMTOM. TomTom Traffic Index – Live congestion statistics and historical data, 2019. URL: https://www.tomtom.com/en{_}gb/traffic-index/.

[6] Europäische Kommission. Transport 2050 : The major challenges, the key measures. *Memo*, 11/197(March):4, 2011. URL: http://europa.eu/rapid/press-release{_}MEMO-11-197{_}ga.htm.

[7] S E H No Mnt. Origin-Destination and Traffic Circulation Study: U.S. 63 Bridge Area . (February), 2009. URL: http://www.dot.state.mn.us/consult/documents/D6SP2515FINALReport.pdf.

[8] Thomas Benz. What does „ Traffic Forecast “ mean ? 2011. URL: https://www.definitions.net/definition/trafficcount.

[9] Nestor Gilbert. 67 Big Data Statistics You Must Learn: 2019 & 2020 Market Share & Data Analysis - Financesonline.com, 2019. URL: https://financesonline.com/big-data-statistics/.

[10] Simon Kemp. Digital 2020: 3.8 billion people use social media - We Are Social UK - Global Socially-Led Creative Agency, 2020. URL: https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media.

[11] Sepideh Kaffash, An Truong Nguyen, and Joe Zhu. Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, 231(December 2019):107868, 2021. URL: https://doi.org/10.1016/j.ijpe.2020.107868, doi:10.1016/j.ijpe.2020.107868.

[12] João Filipe Figueiredo Pereira. Social Media Text Processing and Semantic Analysis for Smart Cities. *arXiv*, (June), 2017. arXiv:1709.03406.

[13] Geotab. What Is Smart Mobility_ _ Geotab. URL: https://www.geotab.com/blog/what-is-smart-mobility/{#}:{~}:text=Smartmobilityreferstousing,walking{%}2Cbiking{%}2Candmore.

[14] Petr Cenek. Intelligent transportation systems. *Komunikacie*, 5(4):10–15, 2003. URL: https://www.intel.com/content/www/us/en/transportation/overview.html, doi:10.1177/1087724x0051001.

[15] Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*, pages 210–215, 2018. doi:10.23919/MIPRO.2018.8400040.

[16] IBM Cloud Education. What is Artificial Intelligence (AI)? | IBM, 2020. URL: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence.

[17] Huimin Lu, Yujie Li, Min Chen, Hyoungseop Kim, and Seiichi Serikawa. Brain Intelligence: Go beyond Artificial Intelligence. *Mobile Networks and Applications*, 23(2):368–375, 2018. arXiv:1706.01040, doi:10.1007/s11036-017-0932-8.

[18] Karthik Sankar. Natural Language Processing Machine Learning. *Artificial Intelligence*, (August 2016):1–13, 2009. URL: http://rutumulkar.com/blog/2016/NLP-ML.

[19] Timothy L Chen, Max Emerling, Gunvant R Chaudhari, Yeshwant R Chillakuru, Youngho Seo, Thienkhai H Vu, and Jae Ho. Domain specific word embeddings for natural language processing in radiology. *Journal of Biomedical Informatics*, 113(August 2020):103665, 2021. URL: https://doi.org/10.1016/j.jbi.2020.103665, doi:10.1016/j.jbi.2020.103665.

[20] Monkeylearn. Guide to Text Classification with Machine Learning. URL: https://monkeylearn.com/text-classification/.

[21] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons.B*, 4(December 2017):51–62, 2017. doi:10.20544/horizons.b.04.1.17.p05.

[22] R. Garreta. A Gentle Guide to Machine Learning, 2015. URL: https://monkeylearn.com/blog/gentle-guide-to-machine-learning/.

[23] Jason Brownlee. What Are Word Embeddings for Text ? pages 1–7, 2017. URL: https://machinelearningmastery.com/what-are-word-embeddings/.

[24] Zellig S. Harris. Distributional Structure. *WORD*, 10(2-3):146–162, aug 1954. URL: https://www.tandfonline.com/action/journalInformation?journalCode=rwrd20, doi:10.1080/00437956.1954.11659520.

[25] Introduction to Word Embedding and Word2Vec | by Dhruvil Karani | Towards Data Science. URL: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa.

[26] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[27] Vivien Cabannes, Alessandro Rudi, and Francis Bach. Fast rates in structured prediction, 2021. arXiv:2102.00760.

[28] Geoffrey E Hinton, Terrence Joseph Sejnowski, et al. *Unsupervised learning: foundations of neural computation*. MIT press, 1999.

[29] IBM Education. What is Unsupervised Learning? | IBM. *Ibm*, (August 2018):1–8, 2020. URL: https://www.ibm.com/cloud/learn/unsupervised-learning.

[30] Farman Ali, Daehan Kwak, Pervez Khan, Shaker El-sappagh, and Amjad Ali. Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowledge-Based Systems*, 174:27–42, 2019. URL: https://doi.org/10.1016/j.knosys.2019.02.033, doi:10.1016/j.knosys.2019.02.033.

[31] Francisco Murços, Tânia Fontes, and Rosaldo J. F. Rossetti. Are bert embeddings able to infer travel patterns from twitter efficiently using a unigram approach? In *2021 IEEE International Smart Cities Conference (ISC2)*, pages 1–7, 2021. doi:10.1109/ISC253183.2021.9562853.

[32] Chris McCormick and Nick Ryan. BERT Fine-Tuning Tutorial with PyTorch · Chris McCormick. pages 1–42, 2019. URL: http://mccormickml.com/2019/07/22/BERT-fine-tuning/https://mccormickml.com/2019/07/22/BERT-fine-tuning/.

[33] Chris McCormick. BERT Word Embeddings Tutorial · Chris McCormick. pages 1–22, 2020. URL: http://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/https://mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/.

[34] Text Analysis Starter Guide: What You Need to Know. URL: https://monkeylearn.com/text-analysis/.

[35] Seyed Mahdi, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems With Applications*, 117:139–147, 2019. URL: https://doi.org/10.1016/j.eswa.2018.08.044, doi:10.1016/j.eswa.2018.08.044.

[36] Monkeylearn. Everything There Is to Know about Sentiment Analysis, 2019. URL: https://monkeylearn.com/sentiment-analysis/.

[37] Veronikha Effendy. Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method. *International Journal on Information and Communication Technology (IJoICT)*, 2(1):57, 2016. doi:10.21108/ijoict.2016.21.85.

[38] Sara Carvalho, Luis Sarmento, and Rosaldo J F Rossetti. Real-Time Sensing of Traffic Information in Twitter Messages. *Proceedings of the IEEE ITSC 2010 Workshop on Artificial Transportation Systems and Simulation (ATSS2010), Madeira Island, Portugal*, pages 19–22, 2010.

[39] Lisa Schweitzer. Planning and social media: A case study of public transit and stigma on twitter. *Journal of the American Planning Association*, 80(3):218–238, 2014. URL: https://doi.org/10.1080/01944363.2014.980439, arXiv:https://doi.org/10.1080/01944363.2014.980439, doi:10.1080/01944363.2014.980439.

[40] A. Candelieri and F. Archetti. Detecting events and sentiment on twitter for improving urban mobility. *CEUR Workshop Proceedings*, 1351:106–115, 2015.

[41] Jose Tomas Mendez, Hans Lobel, Denis Parra, and Juan Carlos Herrera. Using Twitter to Infer User Satisfaction with Public Transport: The Case of Santiago, Chile. *IEEE Access*, 7:60255–60263, 2019. doi:10.1109/ACCESS.2019.2915107.

[42] Mojtaba Maghrebi, Alireza Abbasi, and S. Travis Waller. Transportation application of social media: Travel mode extraction. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pages 1648–1653, 2016. doi:10.1109/ITSC.2016.7795779.

[43] Sonia Anastasia and Indra Budi. Twitter sentiment analysis of online transportation service providers. *2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016*, pages 359–365, 2017. doi:10.1109/ICACSIS.2016.7872807.

[44] N. Nima Haghighi, Xiaoyue Cathy Liu, Ran Wei, Wenwen Li, and Hu Shao. Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. *Public Transport*, 10(2):363–377, 2018. URL: https://doi.org/10.1007/s12469-018-0184-4, doi:10.1007/s12469-018-0184-4.

[45] Daniela Ulloa, Pedro Saleiro, Rosaldo J.F. Rossetti, and Elis Regina Silva. Mining social media for open innovation in transportation systems. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pages 169–174, 2016. arXiv:1610.09894, doi:10.1109/ITSC.2016.7795549.

[46] Bing Qi, Aaron Costin, and Mengda Jia. A framework with efficient extraction and analysis of Twitter data for evaluating public opinions on transportation services. *Travel Behaviour and Society*, 21(May):10–23, 2020. URL: https://doi.org/10.1016/j.tbs.2020.05.005, doi:10.1016/j.tbs.2020.05.005.

[47] Farman Ali, Daehan Kwak, Pervez Khan, S. M.Riazul Islam, Kye Hyun Kim, and K. S. Kwak. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. *Transportation Research Part C: Emerging Technologies*, 77:33–48, 2017. URL: http://dx.doi.org/10.1016/j.trc.2017.01.014, doi:10.1016/j.trc.2017.01.014.

[48] Enrico Steiger, Timothy Ellersiek, and Alexander Zipf. Explorative public transport flow analysis from uncertain social media data. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, GeoCrowd '14, page 1–7, New York, NY, USA, 2014. Association for Computing Machinery. URL: https://doi.org/10.1145/2676440.2676444, doi:10.1145/2676440.2676444.

[49] Craig Collins, Samiul Hasan, and Satish V. Ukkusuri. A novel transit rider satisfaction metric: Rider sentiments measured from online social media data. *Journal of Public Transportation*, 16(2):21–45, jun 2013. URL: https://digitalcommons.usf.edu/jpt/vol16/iss2/2, doi:10.5038/2375-0901.16.2.2.

[50] Shihab Elbagir and Jing Yang. Twitter sentiment analysis using natural language toolkit and Vader sentiment. *Lecture Notes in Engineering and Computer Science*, 2239:12–16, 2019.

[51] Mohammed Abdul-Rahman, Edwin H.W. Chan, Man Sing Wong, Victor E. Irekponor, and Maryam O. Abdul-Rahman. A framework to simplify pre-processing location-based social media big data for sustainable urban planning and management. *Cities*, 109(September

2019):102986, 2021. URL: https://doi.org/10.1016/j.cities.2020.102986, doi:10.1016/j.cities.2020.102986.

[52] Jeffrey H. Kahn, Renée M. Tobin, Audra E. Massey, and Jennifer A. Anderson. Measuring emotional expression with the linguistic inquiry and word count. *The American Journal of Psychology*, 120(2):263–286, 2007. URL: http://www.jstor.org/stable/20445398.

[53] Sentiment Analysis: VADER or TextBlob? | by Bryan White | Towards Data Science. URL: https://towardsdatascience.com/sentiment-analysis-vader-or-textblob-ff25514ac540.

[54] Shreya Ahuja and Gaurav Dubey. Clustering and sentiment analysis on twitter data. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, pages 1–5, 2017. doi:10.1109/TEL-NET.2017.8343568.

[55] Dipanjan Sarkar. Emotion and Sentiment Analysis: A Practitioner's Guide to NLP. *Kdnuggets*, pages 1–6, 2018. URL: https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html.

[56] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation, 2016. arXiv:1609.08144.

[57] Mohd Sanad Zaki Rizvi. What is BERT | BERT For Text Classification, 2019. URL: https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/.

[58] Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch - neptune.ai. URL: https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair.

[59] Narayan CHATURVEDI, Durga TOSHNIWAL, and Manoranjan PARIDA. Twitter to transport: Geo-spatial sentiment analysis of traffic tweets to discover people's feelings for urban transportation issues. *Journal of the Eastern Asia Society for Transportation Studies*, 13:210–220, 2019.

[60] London Population 2021 (Demographics, Maps, Graphs), 2021. URL: https://worldpopulationreview.com/world-cities/london-population.