Universidade Federal de Pernambuco
Centro de Informática
Doutorado Acadêmico em Ciência da Computação

Università di Pisa
Diparimento di Informatica
Dottorato di Ricerca in Informatica

Ph.D. THESIS

Mining human mobility data and social media for smart ride sharing

Vinícius Cezar Monteiro de Lira

SUPERVISORS

Valéria Cesário Times, (UFPE, Brazil)

Chiara Renso, (CNR, Italy)

Rossano Venturini, (UNIPI, Italy)

REFEREE

Stan Matwin

REFEREE

Yannis Theodoridis

“It would not be much of a universe if it wasn’t home to the people you love."
(Stephen Hawking)

Acknowledgements

I would like to express my deep gratitude to Dr. Chiara Renso, Professor Valeria Times, and Professor Rossano Venturini, my research supervisors, for their patient guidance, enthusiastic encouragement and useful critiques of this research work. Their guidance helped me constantly during the time spent in research and writing this thesis. Without doubt, I had the luck to have a really great team of supervisors. I would also like to thank Dr. Raffaele Perego, for his valuable advice and the opportunity to undertake my Ph.D. research within the HPC lab of ISTI-CNR Pisa. It was fantastic to have the opportunity to work in his facilities. My grateful thanks are also extended to Professor Iadh Ounis and Professor Craig McDonald, from the University of Glasgow, for their collaboration in building the predictive models for user attendance using Social Media, and to Dr. Salvo Rinzivillo, from the KDD Lab ISTI-CNR Pisa, who helped me elaborate our novel ride-sharing matching algorithm.

I would also like to extend my thanks to my colleagues of the HPC Lab for the fun and support. The years spent at the HPC Lab were anything but sharing deadlines, hard work, events and travels with such special people. Certainly, all these experiences shared with them made this achievement less hard to be earned. In particular, I would like to individually thank Dr. Patrizio Dazzi, Dr. Massimo Coppola and Dr. Emanuele Carlini for the opportunity to participate in the BASMATI Project developing my skills as researcher.

Moreover, thanks to the external reviewers, Professor Stan Matwin and Professor Yannis Theodoridis, for taking the effort to contribute to this work by reviewing it: I am really honored of having their endorsement. I am also grateful to the UNIPI committee members Professor Anna Monreale and Professor Roberto Grossi for their priceless comments and suggestions to improve my Ph.D. research.

I am especially grateful to my parents for their unconditional love and support in my life. I owe them all I am. I thank as well to all family, who supported me along the completion of the Ph.D. program.

And finally, last but by no means least, I thank all my friends that somehow contributed to my growth during this period.

Abstract

People living in highly-populated cities increasingly suffer an impoverishment of their quality of life due to pollution and traffic congestion problems caused by the huge number of circulating vehicles. Indeed, the reduction the number of circulating vehicles is one of the most difficult challenges in large metropolitan areas. This PhD thesis proposes a research contribution with the final objective of reducing travelling vehicles. This is done towards two different directions: on the one hand, we aim to improve the efficacy of ride sharing systems, creating a larger number of ride possibilities based on the passengers destination activities; on the other hand, we propose a social media analysis method, based on machine learning, to identify transportation demand to an event.

Concerning the first research direction, we investigate a novel approach to boost ride sharing opportunities based, not only on fixed destinations, but also on *alternative destinations* while preserving the intended activity of the user. We observe that in many cases the activity motivating the use of a private car (e.g., going to a shopping mall) can be performed at many different locations (e.g. all the shopping malls in a given area). Our assumption is that, when there is the possibility of sharing a ride, people may accept visiting an alternative destination to fulfill their needs. Based on this idea, We thus propose Activity-Based Ride Matching (ABRM), an algorithm aimed at matching ride requests with ride offers to alternative destinations where the intended activity can still be performed. By analyzing two large mobility datasets, we found that with our approach there is an increase up to 54.69% in ride-sharing opportunities compared to a traditional fixed-destination-oriented approach.

For the second research contribution, we focus on the analysis of social media for inferring the transportation demands for large events such as music festivals and sports games. In this context, we investigate the novel problem of exploiting the content of non-geotagged posts to infer users' attendance of large events. We identified three temporal periods: before, during and after an event. We detail the features used to train the event attendance classifiers on the three temporal periods and report on experiments conducted on two large music festivals in the UK. Our classifiers attained a very high accuracy, with the highest result observed for Creamfields festival ($\sim$91% accuracy to classify users that will participate in the event). Furthermore, we proposed an example of application of our methodology in event-related transportation. This proposed application aims to evaluate the geographic areas with a higher potential demand for transportation services to an event.

Key-words: Ride-sharing, Matching Algorithms, Activity-Based, Social Media, Attendance Prediction.

Resumo

Pessoas que vivem em cidades altamente populosas sofrem cada vez mais com o declínio da qualidade de vida devido à poluição e aos problemas de congestionamento causados pelo enorme número de veículos em circulação. A redução da quantidade de veículos em circulação é de fato um dos mais difíceis desafios em grandes áreas metropolitanas. A presente tese de doutorado propõe uma pesquisa com o objetivo final de reduzir o número de veículos em circulação. Tal objetivo é feito em duas diferentes direções: por um lado, pretendemos melhorar a eficácia dos sistemas de *ride-sharing* aumentando o número de possibilidades de caronas com base na atividade destino dos passageiros; por outro lado, propomos também um método baseado em aprendizagem de máquina e análise de mídia social para identificar demanda de transporte de um evento.

Em relação à primeira contribuição da pesquisa, nós investigamos uma nova abordagem para aumentar o compartilhamento de caronas baseando-se não apenas em destinos fixos, mas também em destinos alternativos enquanto que preservando a atividade pretendida do usuário. Observamos que em muitos casos a atividade que motiva o uso de um carro particular (por exemplo ir a um *shopping center*) pode ser realizada em muitos locais diferentes (por exemplo todos os *shoppings* em uma determinada área). Nossa suposição é que, quando há a possibilidade de compartilhar uma carona, as pessoas podem aceitar visitas a destinos alternativos para satisfazer suas necessidades. Nós propomos o *Activity-Based Ride Matching (ABRM)*, um algoritmo que visa atender às solicitações de caronas usando destinos alternativos onde a atividade pretendida pelo passageiro ainda pode ser executada. Através da análise de dois grande conjuntos de dados de mobilidade, mostramos que nossa abordagem alcança um aumento de até 54,69% nas oportunidades de caronas em comparação com abordagens tradicionais orientadas a destinos fixos.

Para a segunda contribuição nos concentramos na análise de mídias sociais para inferir as demandas de transporte para grandes eventos tais como concertos musicais e eventos esportivos. Investigamos um problema que consiste em explorar o conteúdo de postagens não geolocalizadas para inferir a participação dos usuários em grandes eventos. Nós identificamos três períodos temporais: antes, durante e depois de um evento. Detalhamos as *features* usadas para treinar classificadores capazes de inferir a participação de usuários em um dado evento nos três períodos temporais. Os experimentos foram conduzidos usando postagens em mídias sociais referentes a dois grandes festivais de música no Reino Unido. Nossos classificadores obtiveram alta *accuracy*, com o maior resultado observado para o festival Creamfields (∼91% de *accuracy* para classificar os usuários que participarão do evento). Propusemos também uma aplicação de nosso método que visa avaliar as áreas geográficas com maior potencial de demanda por serviços de transporte para um evento.

Palavras-chaves: Compartilhamento de caronas, Algoritmos de Matching, Activity-Based, Mídias Sociais, Predição de participação.

Sommario

Le persone che vivono in città densamente popolate subiscono sempre più un impoverimento delle loro qualità della vita a causa dell'inquinamento e dei problemi di congestione del traffico causati dall'enorme numero di veicoli circolanti. La riduzione dei veicoli circolanti è una delle sfide più difficili nelle grandi aree metropolitane. Questa tesi di dottorato propone un contributo di ricerca con l'obiettivo finale di ridurre i numeri di veicoli in viaggio. Questo eśtato sviluppato verso due direzioni: da un lato, vogliamo migliorare l'efficacia dei sistemi di *ride sharing*, aumentando la possibilità di ricevere e dare passaggi in base alla attività di destinazione dei passeggeri. D'altra parte, vogliamo proporre un metodo basato sul *machine learning* e analisi dei *social media*, per identificare demanda de transporte a un evento.

Per quanto riguarda il primo contributo di ricerca, abbiamo studiato un nuovo approccio per aumentare la condivisione dei passagi non solo su destinazioni fisse, ma anche su destinazioni alternative preservando l'attività prevista dall'utente. Osserviamo infatti che in molti casi l'attività che motiva l'uso di un'auto privata (ad es. andare in un centro commerciale) può essere eseguito in molti luoghi diversi (ad esempio tutti i centri commerciali in una determinata area). La nostra ipotesi è che, quando c'è la possibilità di condividere un passaggio, le persone possono accettare di visitare una destinazione alternativa per soddisfare i loro bisogni. Basato su questa idea, proponiamo *Activity-Based Ride Matching (ABRM)*, un algoritmo che mira a soddisfare le richieste di *carpool* utilizzando destinazioni alternative, dove l'attività desiderata dal passeggero può ancora essere eseguita. Attraverso l'analisi di due grandi insiemi di dati di mobilità, mostriamo che il nostro approccio raggiunge un aumento fino al 54,69% nelle opportunità di condivisione di car pooling rispetto agli approcci tradizionali rivolti a destinazioni fisse.

Per il secondo contributo della ricerca ci concentriamo sull'analisi dei social media per inferire le richieste di trasporto verso grandi eventi come concerti musicali e giochi sportivi. In questo contesto, indaghiamo sul nuovo problema dello sfruttamento del contenuto di *non geotagged post* per inferire la presenza di utenti a grandi eventi. Abbiamo identificato tre periodi temporali: prima, durante e dopo un evento. Descriviamo in dettaglio le caratteristiche utilizzate per addestrare i classificatori per inferire la partecipazione all'evento sui tre periodi temporali. Riportiamo gli esperimenti condotti su due grandi festival musicali nel Regno Unito. I nostri classificatori raggiungono una alta *accuracy*, con il risultato più alto osservato per il festival Creamfields ($\sim$91% di *accuracy* per classificare gli utenti che parteciperanno all'evento). Inoltre, abbiamo proposto un'applicazione della nostra metodologia che ha come scopo valutare le aree geografiche con il maggior potenziale di domanda di servizi di trasporto per un evento.

Parole chiave: condivisione di passaggi, algoritmo di matching, condivisione di passaggi basato su attivitá, reti sociali, predizione di parteipazione.

Contents

List of Figures

List of Tables

1 Introduction

In large cities, mobility is one of the most critical issues for a proper functioning of urban areas. Inappropriate transportation infrastructures and services produce several negative impact on the quality of life, such as increasing pollution and traffic congestion.

Many studies have shown the future of mobility to be a major focus of research for the next decade and beyond (URRY, 2016; SPICKERMANN; GRIENITZ; HEIKO, 2014; WEGENER, 2013). A recent study by TomTom (TOMTOM, 2017) shows how the traffic situation is often critical and heavily affects people mobility in cities like Istanbul (Turkey), Mexico City (Mexico), Rio de Janeiro (Brazil) and Moscow (Russia). Istanbul is the worst city in this unenviable ranking, with a daily average delay of 29 minutes for a 30 minutes commute. Moreover, traffic is not just a discomfort for drivers, but it can also harm the environment and brings negative consequences for the economy. In the United States, transportation studies report the annual cost of congestion at $160 billion, which includes 7 billion hours of time lost to sitting in traffic and an extra 3 billion gallons of fuel burned (ALONSO-MORA et al., 2017).

A multidisciplinary topic of research that addresses issues raised by the uncontrolled growth of urban centers is the development of Smart Cities (FARKAS et al., 2015; TOWNSEND, 2014; ALKANDARI; ALNASHEET; ALSHEKHLY, 2012). In (BENEVOLO; DAMERI; D'AURIA, 2016), the authors define Smart City as a set of urban strategies using technology to improve the quality of life in urban spaces, by improving the environmental quality and delivering better services to citizens. Specifically, when dealing with mobility issues in large urban centers, Smart Cities rely on one of its most promising pillar: the Smart Mobility services (BENEVOLO; DAMERI; D'AURIA, 2016). A Smart Mobility environment offers a whole ecosystem of solutions for reducing congestion and fostering faster, greener, and cheaper transportation options. This ecosystem is composed of several solutions using advanced technology aimed at providing alternative sustainable and integrated options of transportation (PFRIEMER, 2017; MASRI et al., 2017). Thus, in an increasingly connected and populated society, the Smart Mobility Ecosystem provides access not only to the full range of public transport options but also to all the add-ons, such as car-sharing, bike-sharing and ride sharing services, such as Uber and Lyft, currently growing globally (FLÜGGE, 2017; CHIANG et al., 2018; JIANG et al., 2018).

The focus of this thesis is on improving ride sharing systems as a possible solution to reduce the number of circulating vehicles. Ride sharing consists of the sharing of a vehicle by two (or more) persons who move along similar itineraries and time schedules. Thus, it is designed around individual mobility needs and it relies on a simple resource: the empty seats in cars (STIGLIC et al., 2016; MASRI; ZEITOUNI; KEDAD, 2017; MA; WOLFSON, 2013). In (FURUHATA et al., 2013), the authors identify three major challenges for ride sharing

systems: design of attractive matching mechanisms, proper ride arrangement, and building of trust among unknown travelers in online systems. Hence, most of these systems deal with an optimization problem: how to determine the routes and schedules of the vehicles, including how to assign the demand of rides to drivers, solving conflicting objectives, such as maximizing the number of participants, minimizing the trip cost or minimizing the passenger inconvenience (AGATZ et al., 2012).

Indeed, every day millions of people drive alone in parallel with neighbors who very often are driving to similar locations. A recent study by the U.S. Census Bureau (FLORIDA, 2015) has shown that three out of four Americans (76.4%) report driving to work alone. Almost ten percent (9.4%) carpool, though this figure has actually declined significantly from a high of nearly 20 percent in 1980. These empty seats in cars represent a huge waste of resource in transportation system, but potentially also a huge opportunity for improvement. In this scenario, many works in literature have proposed ride sharing solutions to avoid single vehicle occupancy trips and attract enough participants to achieve a satisfactory mass of users (STIGLIC et al., 2016; GEISBERGER et al., 2010; FURUHATA et al., 2013; TRASARTI; GIANNOTTI; NANNI, 2011).

The prearrangement process to match the supply and demand is a key characteristic of ride sharing. In this context, (FURUHATA et al., 2013) classify service providers into two types: ***matching agencies*** and ***transportation service operators***. With *matching agencies* we refer to service providers which facilitate ride sharing services by matching between individual car drivers and passengers. Example of these kind of services are the popular known BlaBlaCar [1], Lyft [2] and Uber [3]. These services exploit different kinds of *matching algorithms* to find the best allocation of ride offers to ride requests. These algorithms are typically based on the spatial and temporal aspects of the rides.

With *transportation service operators* we refer to organizations which provide ride sharing services identifying and supplying the demand of rides with their own vehicles and drivers, such as airport and hotel shuttles. Current researches in this context include both *routing algorithms* for optimal supply of the demand of rides and *inference of rides demand* to support transportation planning.

This thesis provides a research contribution in both these scenarios. We propose a ride matching algorithm for alternative destinations and a classification method to identify potential users for a transportation service towards large events.

The present chapter discusses the motivations that guided this PhD thesis in Section 1.1, introduces the objectives of the thesis in Section 1.2 and illustrates the structure of the thesis in Section 1.3.

[1] www.blablacar.com
[2] www.lyft.com
[3] www.uber.com

1.1 Motivation

As we have introduced in the previous section, this thesis proposes a twofold novel research contribution: improving ride matching algorithms with alternative destinations and identifying potential users for transportation services by exploring the content of posts on microblogging platforms. They both have the final objective of reducing the number of circulating vehicles.

We discuss the specific motivations and approach for each research line in the following.

Improving the efficacy of ride matching algorithms. Ride matching algorithms are used to find feasible associations between driver and passengers (FURUHATA et al., 2013). Basically, the drivers offer rides using their vehicles and, in turn, the riders seek for available rides to reach their desired destination. A feasible matching must satisfy certain spatial-temporal constraints such as: passenger's pick-up time, pick-up location and destination, and drivers' route and departure time (FURUHATA et al., 2013). However, when there are few ride offers available, it might be difficult to find feasible matches (WANG et al., 2016; STIGLIC et al., 2016). Therefore several methods to enhance the number of possible matchings have to be developed in the literature. A largely adopted solution to improve ride-sharing opportunities consists in relaxing the spatio-temporal constraints (STIGLIC et al., 2016; GEISBERGER et al., 2010; FURUHATA et al., 2013; AGATZ et al., 2012) such as delay and walking distance tolerance of the passengers or drivers' route detours. However, these approaches focus only on spatial and temporal constraints for the matching between ride offers and ride requests, without exploring a valuable dimension: the semantics behind the ride requests. Here, with semantic we refer to the *activity* that is going to be performed by the passenger at the trip destination. Some research proposals related to human movement behavior have investigated important factors in trip-making choices and identified that the activity to be performed at the destination plays a crucial role (CHEN et al., 2016; PELEKIS; THEODORIDIS, 2014; KITAMURA, 1988).

Recent studies of human mobility highlight the significant potential for continued and future uptake of sustainable forms of urban mobility (RODE et al., 2015; STEG; VLEK, 2009), while at the same time investigating the tendency to be regular or not in choosing the places where to perform some activities (LIRA et al., 2014; WU; LI, 2016).

Based on these observations, we postulate that changing mobility habits can be rewarding under some aspects, like avoiding traffic congestion or saving money. As a consequence, people can accept to change their trip destination to some alternative location, when there is the possibility of sharing a ride, thus saving time and/or money and enforcing their pro-environment behavior (STEG; VLEK, 2009).

We observe that, for some activities, such as grocery shopping or ATM services, people often have a set of alternative places that they use for the same purpose. Such findings, in the context of ride sharing, open space for a novel investigation toward the massification

of ride sharing system by exploring a new input dimension: *"Can the usage of ride sharing system be boosted by exploiting alternative destinations based on the intended activity of the passenger"?*

Identifying transportation demands for large events. Large events like important music concerts, religious celebrations or sports matches, attract thousands of participants. These events cause the movement of thousands of people to a specific location at a given time. In general, a large event requires careful transportation planning to facilitate the attendees' arrival at the event location from their diverse departure locations. For this reason, often the event organizers provide the participants with dedicated transportation services (e.g. bus, carpooling, shuttle) to supply the demand for rides. However, for large events involving a massive number of attendees, the identification of this demand of rides might not be a straightforward process, requiring new methods to devise them (SINNOTT; CHEN, 2016a). As a second contribution of this thesis, we investigate: *"How can we identify the actual attendance to a given event for a potential passenger of a transportation service?"*.

Our investigation starts from observing that large events are well reflected in social media (e.g. Facebook, Foursquare, Instagram, Twitter). Indeed, people can connect to "the event" by expressing through *posts* their feelings, experiences or opinions about the event well in advance with respect to its planned date. Moreover, social media have been often exploited to extract valuable information concerning human dynamics and behaviors (CHO; MYERS; LESKOVEC, 2011; CESARIO et al., 2016; QUERCIA et al., 2010). Due to this vast applicability, social media analytics is a fast growing research area, aimed at extracting useful information from user generated data (ALBUQUERQUE et al., 2016; SOKOLOVA et al., 2016; GAL-TZUR et al., 2014a).

Thus, given the attention to popular events reflected in social media, this thesis addresses a new challenging problem: *"Is it possible to infer from social media posts the actual attendance of the media user to the cited event?"*. If we could classify user posts discussing an event on the basis of the actual attendance of the user to the event, we could enable or enhance several practical applications not only in ride sharing, but also for example, of targeted advertising and mobility management. Furthermore, from this analysis, we want to derive the key point of our investigation: *"Is it possible to infer past, current and future user attendance to large events through posts on social media to forecast the demand of rides?"*. By inferring the future attendance, we can predict the users who will attend the event and potentially identify the ride demands to reach the event's location. While, by inferring current and past attendance, we aimed at understanding who were the users who moved to the location of the event. These two latter subsets can support future transportation planning for the next editions of the event.

The simplest way of inferring the presence of users at events is considering the location associated with their media posts: the geotag, or "check-in", indicates the user presence at

the time of the event at the event location. We observe, however, that this approach suffers from two drawbacks. The first drawback is that a low number of social media users enable geotagging of their posts (in Twitter the percentage of geotagged posts is reported at about 2% (LEETARU et al., 2013; SLOAN et al., 2013)). Geolocation information is geographically accurate, but geotagged media posts represents very sparse data source. Using sparse data to learn attendance prediction classifiers becomes extremely difficult and produces models with limited applicability. The second drawback of using only geolocated posts is that they represent the actual presence of the user at the event but not the *intention* of the user to participate in the event. Indeed, we aim to infer the user participation to not only the event before, but also after the event takes place. The early knowledge of the possible user attendance can be extremely useful for enabling innovative services and applications in the fields of transportation planning and crowd safety management (KAISER et al., 2017).

The next section of this chapter details the objectives of this thesis.

1.2 Thesis Objectives

The aim of the present thesis is two-fold. Firstly, we aimed at investigating how ride matching algorithms can be improved in efficacy by taking advantage of passenger flexibility. For this purpose, we propose Activity-Based Ride Matching (ABRM), an algorithm aimed at matching ride requests with ride offers possibly reaching alternative destinations where the intended user activity can be performed. By exploiting the knowledge of the activity motivating ride requests, ABRM can boost users' mobility demands by means of existing ride offers. The approach proposed is completely orthogonal and possibly complementary to popular ride sharing services like BlaBlaCar[4], UberPool[5] and Lyft Line[6]. Indeed, providing the user with activity-based ride options could enable novel business strategies to be incorporated in these services. For example, the service could support user's flexibility and increase her engagement by proposing the most convenient rides to alternative destinations where the intended activity can be performed.

Secondly, we aim at deriving the demand of rides for large events through the use of social media. In order to do that, we want to overcome the issues of inferring the actual attendance of users based on geo-tagged post by taking a novel strategy which relies on the content of *non-geotagged posts* only, without considering any spatial features. Moreover, we perform our event attendance classification by distinguishing three temporal intervals identifying when the posts have been shared on social media: *before*, *during* or *after* the event. The posts shared before the event may express the interest of the users in the upcoming event and their intention to attend, or their regrets for not being able to attend

[4] www.blablacar.com
[5] www.uber.com/ride/uberpool/
[6] www.lyft.com/rider

it. During the event, people may express their feelings about the event, may report issues with the provided services or may also share photos and videos. After the event, users may report feelings and comments on their experience at the event. The analysis of posts shared before the event acts as a prediction of the users' actual attendance, during the event reflects the actual participation of users at the event, while after the event it gives a view of past attendance. To solve this challenge, we use machine learning techniques to explore features extracted from event-related posts to identify attendees of large events.

Thus, the novel contributions of this thesis can be shown as follows:

1. In the context of ride sharing matching algorithms, we propose an algorithm for boosting ride sharing usage by exploiting alternative destinations based on the user intended activities.

    a) We propose a novel ride sharing matching algorithm that considers alternative destinations to improve efficacy;

    b) We define a number of ride matching features for the evaluation of the qualities of the rides retrieved by the proposed matching algorithm;

    c) We design and run extensive experiments exploiting mobility datasets representing real user activities and mobility demands for these activities at a large scale;

    d) We analyze and discuss in depth the potential environmental impact of the proposed solution compared to state-of-the-art techniques.

2. In the context of ride sharing service operators, we propose a method to estimate the demand of rides to large event by inferring event-attendance through posts on social media. The idea is to understand transportation demand by using machine learning to infer the past, present and future users' attendance based on the content of the event-related posts.

    a) We propose a classification task for inferring users' attendance to events by using posts on social media;

    b) We define relevant post-based features for the classification task;

    c) We design and run extensive experiments for assessing the accuracy performance of our classifiers;

    d) We apply the proposed approach in a real world application for the identification of the demand of rides, using a real large event.

The next section presents the structure of the remaining chapters of this thesis.

1.3 Structure of the Thesis

The rest of this document is organized as follows:

- **Chapter 2** describes the basic concepts of this thesis including the main characteristics about ride sharing systems.
- **Chapter 3** presents the results of a bibliographic survey. The focus is to review the main strategies used by the matching algorithms to increase the amount of matching in a ride sharing system and, subsequently, to optimize ride assignments between drivers and passengers. In this chapter, we also present how large events in social media have been extensively studied for inferring event attendance.
- **Chapter 4** introduces the first research contribution of the thesis, the Activity-Based Ride Matching (ABRM) algorithm. This algorithm takes into account the intended activity of the passenger to suggest possible rides to alternative destinations. Additionally, a rank model is proposed to sort candidate rides according to the ride request requirements. Finally, a demo using ABRM is shortly presented.
- **Chapter 5** introduces the second research contribution of the thesis, a machine learning approach using social media data to infer event attendance to large events. We also introduce and discuss a real world application of the proposed approach that estimates demand of rides to a music festival in UK.
- **Chapter 6** closes this thesis by providing a summary of the contributions and a discussion about their limitations. This chapter also presents some possible future research directions exploiting our results. Finally, it exhibits the list of scientific publications achieved during this PhD research.

1.4 Conclusions

This chapter presents the contextualization, the motivations, the objectives and structure of this thesis. The following chapter presents the basic concepts in ride-sharing systems that are needed for the understanding of this thesis.

2 Basic Concepts

This chapter presents the basic concepts needed to facilitate the comprehension of this thesis proposal. Section 2.1 highlights the differences among Car sharing, Carpooling and Ride sharing. Section 2.2 discusses the characteristic of Ride Sharing Systems, while the Section 2.3 introduces two types of Ride Sharing Providers: Matching Agencies and Transportation Operators. Finally, the Section 2.4 concludes this chapter.

2.1 Car sharing vs Carpooling vs Ride sharing

Often the concepts of "car sharing", "carpooling" and "ride sharing" cause misunderstanding among them. These concepts are the result of evolution over time and of the progressive importance that the collaborative economy is taking on. However, all these three services are seen as emerging alternative transportation modes. They are eco-friendly and sustainable as they enable people to save time, share resource costs, reduce emission and traffic congestion (GALLAND et al., 2014). Then, here we clarify these possibilities, outlining the main features of each services:

- **Car sharing.** Consists of the rental of a car owned by third parties, generally short-term, often by minute or hour, in urban areas. The same car is made available to more drivers who use it individually for a limited time (MARTIN; SHAHEEN; LIDICKER, 2010; WEIKL; BOGENBERGER, 2013). The renting organization may be a commercial business.
- **Carpooling.** It is intended that sharing of the trip does not provide a gain for the driver but only a sharing of costs, or a courier transport activity. It can also be seen as a particular kind of ride sharing, where one of the users shares her own car to offer rides to other passengers (CORREIA; VIEGAS, 2011; GALLAND et al., 2014).
- **Ride sharing.** Refers in general to the activity of sharing car-rides, also in order to produce a profit (in this case called "rides on demand") (FURUHATA et al., 2013).

Since in this thesis we do not focus on the business models of these approaches, but on the sustainable aspects, we consider both terms Carpooling and Ride sharing as similar from the practical point of view of the sharing of rides between drivers and passengers.

2.2 Ride Sharing

Ride sharing is not new. In 1942, during World War II, the U.S. government encouraged ride sharing arrangements in workplaces when no other transportation options were avail-

able in order to save rubber (CHAN; SHAHEEN, 2012). In the 1970s, the oil crisis and spike in gasoline prices encouraged another period of ride sharing. However, today's ride sharing revolution was made possible by the development of GPS, smart phone technology, and electronic payments (AGATZ et al., 2012).

The term ride sharing literally means to share a ride. Ride sharing is a mode of transportation consisting of two (or more) persons sharing a vehicle to move along similar itineraries and time schedules. While public transit options, like the bus, may be cheap, they are plagued by inconvenient schedules, stops and unexpected rider problems, a ride sharing system can be seem as a system that can combine the flexibility, the convenience and the speed of private cars with reduced cost (STIGLIC et al., 2016; FURUHATA et al., 2013; MA; WOLFSON, 2013). Furthermore, from a business point of view, the idea of sharing car rides is part of a global trend, called "Sharing economy", that follows the principle of sharing goods, experiences and knowledge and collaborate to get the best from the shared resources. A mindset shift is needed to get into the idea that sharing can bring more value than ownership since everyone gets more benefit (GARGIULO et al., 2015).

Indeed, carpooling or ride sharing has been an effective way for people with similar schedules and work locations to help meet the demands of one another (CAULFIELD, 2009; FURUHATA et al., 2013). Thus, ride sharing offers advantages for the participants (both drivers and passengers), to the society, and to the environment in addition to saving travel cost, reducing travel time, mitigating traffic congestions, conserving fuel, and reducing air pollution (JIANG et al., 2018).

A successful ride sharing requires coordination with respect to itineraries that include the specification of the passenger's pick-up and drop-off (AGATZ et al., 2011; FURUHATA et al., 2013). This coordination can, in addition, take into account other issues, such as travel cost, compensation for alternative ride provision, gender, and reputation of drivers and passengers. The prearrangement can start when ride requests or offers are submitted to the service providers, which then aim to match the supply and demand for rides.

2.3 Types of Ride Sharing Service Providers

(DAILEY; LOSEFF; MEYERS, 1999) defines a trip as a single instance of travel from one geographic position to another. In this context, a ride can be seen as joint-trip of at least two travelers (one driver and at least one passenger or in a near future at least two passengers been transported by a self-driving car) that share a vehicle. Each participant has a demand for her trip consisting of the origin and the destination. It is not relevant if both participants share the same trip motivation, but necessarily they have similar itineraries and schedules.

Peculiar characteristics on the way how the rides are arranged rule the distinctions among the ride sharing systems. The arrangement of a ride is strictly associated to the

mobility purpose of the travelers. The demand might be to travel long distances, for example when crossing cities, or even short distance, moving from home to a bar. In some systems, rides can be arranged on demand or on real time requiring a short-term agreement that can take just few minutes, in other cases instead, it might involve long-term agreement, usually pre-established days or hours before. For these reasons, several ride sharing service providers operate for different kinds of mobility demands and, as consequence, there is not a unique business strategy ruling these systems.

As proposed by (FURUHATA et al., 2013), we divide ride sharing service providers into two groups:

- **Transport Service Operators:** operate ride sharing services using their own vehicles and drivers.
- **Matching Agencies:** facilitate ride sharing services by performing matching between individual car drivers and passengers.

Ride sharing takes on different characteristics when it is run by service operators and when is it coordinated by matching agencies. These groups are futher discussed in subsections 2.3.1 and 2.3.2.

2.3.1 Transportation Service Operators

This group of systems is more traditional. Representative examples of service operators are vanpooling and airport shuttle transportation services. Typically, they accept requests from passengers and assign these ride requests to vehicles that they operate. Here, the driver is often an employ of the company providing the transportation service (FURUHATA et al., 2013).

Some service operators specify either a fixed and common pick-up or drop-off location for all the passengers, while others allow passengers to choose both. For example, for the airport shuttle, travelers may be picked-up from different hotels to head to a common destination, the airport. Pick-up and drop-off times can sometimes require some amount of slack time. One challenge faced by these systems is to solve the Vehicle Routing Problem (VRP) for the determination of the optimal set of routes to be performed by a fleet of vehicles to serve a given set of passengers (TOTH; VIGO, 2002). In general, the ride demand must be known in advance to guarantee a proper operation and a optimal routing plan.

Figure 1 shows a transportation service operation schema. We notice that the transportation demands and the vehicles are the inputs for the logistic problem solver, while optimal vehicle routes are the output of this operation.

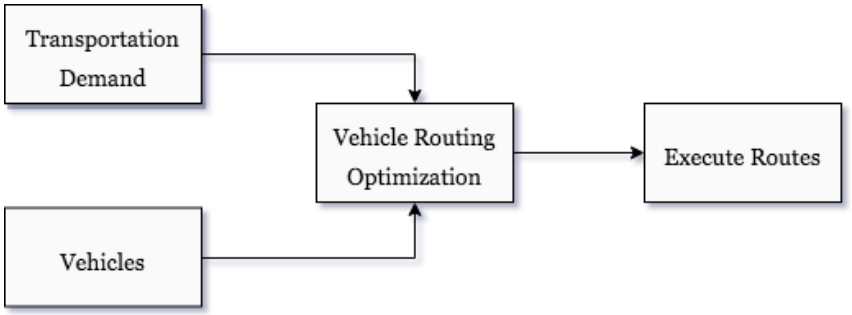


Figure 1 – A *transportation service* operation schema

2.3.2 Matching Agencies

In contrast, matching agencies focus on ride-matching services between individual car drivers and passengers. Unlike service operators, matching agencies do not provide vehicles and drivers. Instead, individual drivers have their own trip plans and provide unoccupied seats for passengers to share their travel expenses (FURUHATA et al., 2013). Furthermore, more recent systems such as Uber and Lift, allow part-time drivers to use their own car to find riders and get paid for the rides.

Matching agencies use ride offers and ride requests received from drivers and passengers, respectively, to find suitable ride sharing matches. A ride offer represents a vehicle trip with available seats to other passengers (riders). This could be for example private cars performing a regular trip going home from work at 5pm of working days or performing a trip between two cities, or could be even a shared taxi serving a ride. On the other hand, the ride request represents the intention of the rider(passenger) to move from a initial point to destination by means of a ride. This could be for example, a rider at home intending to move to a shopping center at 6 pm or even a rider looking for rides toward a specific city.

For a matching, all of the participants must agree on the costs and schedules, which depends on the routes used, including the pick-up and drop-off locations of passengers. In general, the ride arrangement is performed in a short time taking minutes or even seconds for the agreement between drivers and passengers. However, the ride matches involving long trips as border-crossing might require a longer-term agreement. Matching agencies use ride sharing matching algorithms to find feasible ride matchings. The value of a ride matching algorithm depends on how efficiently and effectively suitable matches can be found (FURUHATA et al., 2013).

A ride sharing matching algorithm can be conceptually described by the schema in Figure 2. It takes as input two sets: *ride offers* of users willing to share their trips, and *ride requests* of users searching for a lift to a destination. A *Ride Matching* phase retrieves a subset of ride offers that may supply a ride to a given ride request. In this phase, the system finds all matchings that are compatible with spatial and temporal constraints specified by the user. After the possible matchings have been identified, the *Ride Allocation* phase

uses some criteria to allocate ride requests to offers, limited by the vehicle capacity and based on the matchings. Usually an optimization strategy is applied at this phase aimed at optimizing the allocation between driver and passenger according to a specific criteria such as maximization of the number of participants or minimization of the trip cost for the participants involved. Since it is desirable that both passengers and drivers find a ride matching that best meets individual preferences, the *User Acceptance* phase checks which suggestions have been accepted by the users and, in case of rejection, looks for an alternative allocation repeating the previous step of Ride Allocation. To maximize the probability of successfully matching, it is crucial to provide the Ride Allocation phase with a large set of candidates.

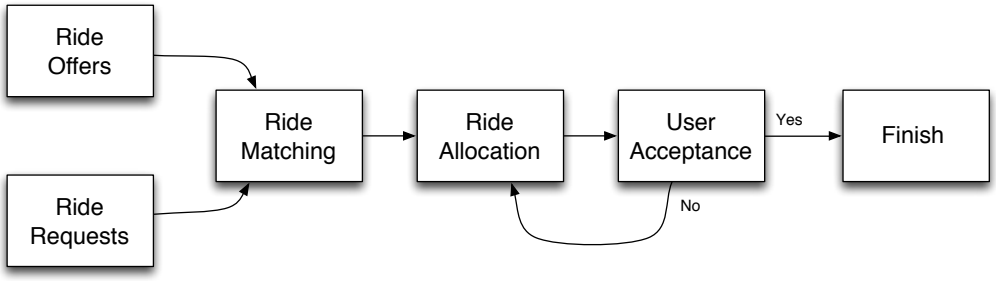


Figure 2 – A ride *matching* algorithm conceptual schema

2.3.3 Ride Matching Problem (RMP)

The central element in a matching agency system is the ride matching. For this reason, we need a clear comprehension of what is the ride matching problem and what are the constraints involved.

The ride matching problem consists in finding feasible matchings between ride offers and ride requests under certain constraints. Furthermore, given the different possible combinations of matchings between these elements, the ride-matching problem can be treated as a combinatorial optimization problem (TEODOROVIĆ; DELL'ORCO, 2008). The following are potential objective functions of this optimization problem: (a) minimize the total vehicle distance traveled; (b) minimize the total delay; (c) make vehicle utilization relatively equal; and/or (d) minimize travel costs.

For the retrieval of feasible ride matchings, three important constraints must be considered: spatial, temporal and capacity constraints. In principle, a feasible ride matching must satisfy all these three constraints. We discuss them in the following subsections.

2.3.3.1 Spatial Constraints

In ride matching systems, a ride request can be matched with a driver having a similar itinerary that includes both the requested pick-up location and to the intended destination. The pick-up location represents the meeting point between the driver and one or

more passengers, it is basically the point where the ride starts. The drop-off location instead represents the end of the ride in a point nearly the passenger's intended destination. In ride matching, as requirement, both the pick-up and the drop-off locations must be agreed.

Furuhata et. al. in (FURUHATA et al., 2013) classified ride sharing matchings according to these two positional elements. Here, we extend their proposed classification, aligning it with the current state of the art and classifying ride sharing matches into four spatial patterns. In the following, we describe these patterns for the single passenger case. To facilitate the comprehension, we present some symbols and notations of the elements of a ride. Let us denote g as a passenger and R_i as an original route of a driver i. The passenger g has his origin and destination locations defined by O and V, respectively. We denote P as a pick-up location and D as a drop-off location. Consider also the functions *detour* and *walk*. The function *detour* performs a detour in the original route R_i. Instead, the function *walk* represents the walking of the passenger p from one location A to a location B.

- **Inclusive Matching**: Both the origin O and the destination V of passenger g is on the way of an original route R. Figure 3 illustrates an inclusive matching. Thus, $O = P$, $D = V$ and $P, D \in R$.

Figure 3 – Inclusive matching example

- **Partial Matching**: Either the origin O or the destination V or both of the passenger g is not on the way of the route R. However, by walking either to pick-up point $walk(O, P)$ or to drop-off point $walk(D, V)$ or to both, then both the pick-up location P and drop-off location D of passenger g are on the way of an original route R. Figure 4 illustrates a partial matching. Thus, $O \neq P$, $D \neq V$, and $P, D \in R$.

- **Detour Matching**: Either the pick-up location P or drop-off location D or both are not on the way of an original route R. However, by taking a detour, $detour(R)$ covers one or both the pick-up and drop-off locations. In general, the detour of any participant is calculated as the ratio between the minimal additional distance necessary for a driver to match the ride request and the original route R of the driver. Figure 5 illustrates a detour matching. Thus, $O = P$, $D = V$ and $P, D \in detour(R)$.

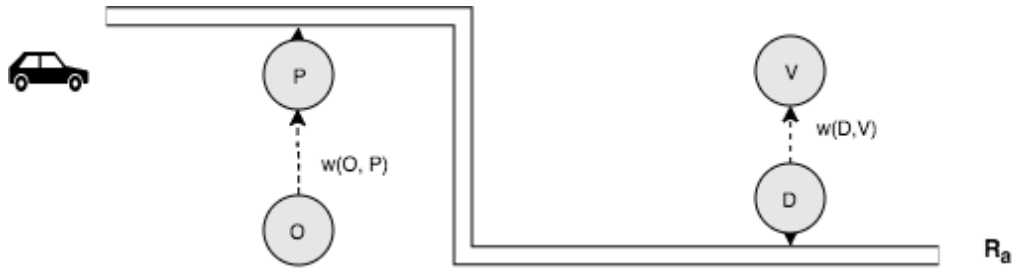


Figure 4 – Partial matching example

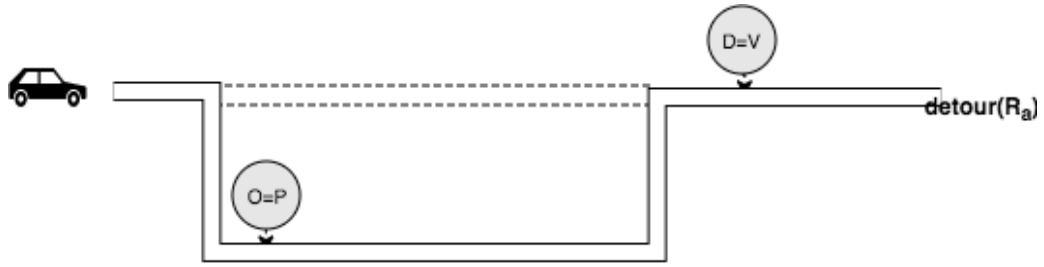


Figure 5 – Detour matching example

- **Multi-hop**: The driver a picks-up the passenger g at the location P^a that is on the way of R_a. The first drop-off location D^a of the ride is not at the destination location V, but it is close to another route R_b performed by the driver b. Then, the driver b picks-up the passenger g at the location P^b and drops the passenger g at his destination D. Thus, $O = P^a$; $D^a = P^b$; $D^b = V$; $P^a, D^a \in R_a$; $V \notin R_a$; $P^b, D^b \in R_b$; $P^a \notin R_b$. This classification of ride-sharing system can also consider that the passenger can walk a short distance to reach the pick-up points.

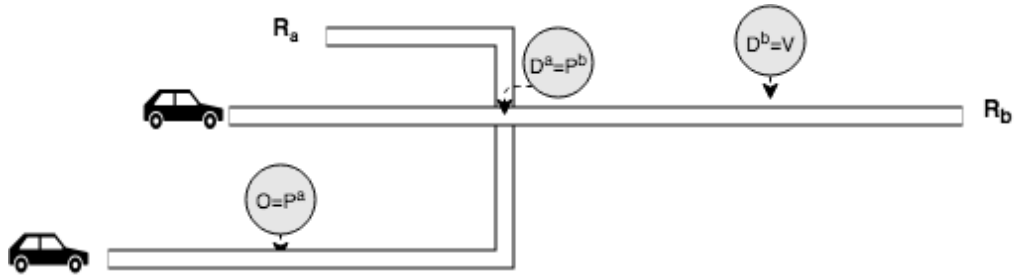


Figure 6 – A pool *ride* Multi-hop

It is important to mention that traditional matching agencies do matching between ride offers and ride requests via proximity rather than the exact locations. Note also that, usually, when a partial matching occurs, the pick-up and drop-off locations are either given input as if their origins and destinations are located on major streets or determined by negotiations. In addition, passengers need to find an alternative transportation method to complete their trips or just walk.

The path between the pick-up P and the drop-off D of a passenger defines the *ride length*. In many matching agencies, the *ride length* is the main criteria for the definition

of the cost of the trip that will be shared among the participants of the ride.

2.3.3.2 Temporal Constraints

A ride matching requires also an agreement between the driver and the passengers with respect to the schedule of the pick-up and the drop-off times. Basically, the **pick-up time** is the time at which the passenger joins into the ride and, in turn, the **drop-off time** is the time at which the passenger leaves the ride. Thus, the **ride duration** is defined by the difference between the drop-off and the pick-up times.

This constraint could limit the amount of matchings since it is very unlikely that a ride request finds a ride offer at the exact time of the request. A common approach to overcome this problem is to relax the temporal constraint by using time windows either for departure or for arrival, or both. Thus, in the context of ride sharing, the time window represents acceptable delays or anticipation in time.

To facilitate the understanding of these concepts, we introduce the following notations: let P_e and P_l be the *earliest and the latest pick-up time*. Similarly, let D_e and D_l be the *earliest and the latest drop-off time*. These notations are illustrated in Figure 7.

The **pick-up time window** is the temporal interval during which a passenger can be join a ride, comprising the range $[P_e, P_l]$. The pick-up time window is defined either by the passenger or by a ride sharing system's policy that may inform a maximum tolerance delay for the driver to pick-up the passenger. For example, if the pickup time is 10:00 AM and the passenger has a maximum delay tolerance of 15 minutes, then the pickup time window is between 10:00 AM and 10:15 AM. Of course, the wider the time-window, the wider the possibility for the passenger to find a ride.

Similarly, the **drop-off time window** comprises the temporal range $[D_e, D_l]$, defining a minimum and a maximum time at which the passenger may arrive at the drop-off location. The use of drop-off time windows is less common or not used as constraint for the ride matching problem.

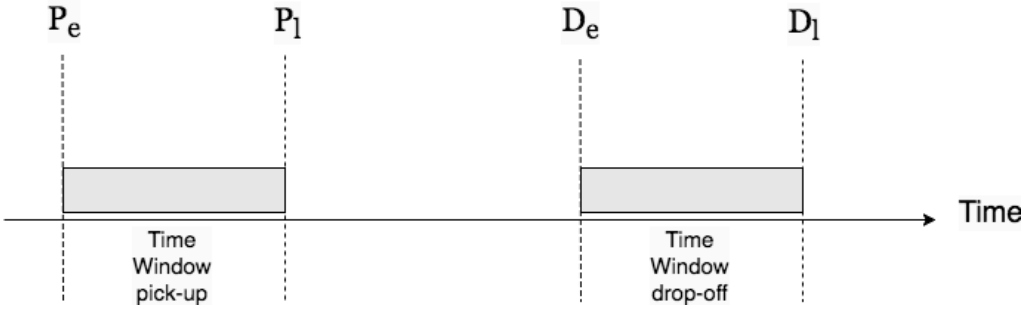


Figure 7 – Pick-up and drop-off time windows

2.3.3.3 Capacity constraints

The demand of a ride is defined by the number of passengers that will be delivered from the pick-up to the point of drop-off. This can be specified on the request, indicating the number of passengers needing a ride. Of course, in a ride, the number of participants that simultaneously join the ride cannot be higher than the **vehicle's capacity**. Some ride-matching algorithms might accept the pool of the vehicle, picking up more passengers along the trip in different pick-up locations.

2.4 Conclusions

This chapter presents the theoretical basis necessary for understanding this thesis, in which concepts and characteristics related to ride sharing systems and the matching algorithms for ride sharing were described.

The following chapter presents the results of a bibliographic survey discussing the most recent approaches that address the problem of ride matching. Furthermore, we discuss some works that study event attendance through the use of social media.

3 Related Work

In this chapter, we discuss several works that have common purposes with this thesis. Section 3.1 details several studies that address the problem of matching between ride offers and ride requests. In turn, Section 3.2 discusses recent work in literature concerning studies of user attendance in large event through the use of social media. Finally, Section 3.3 presents the final considerations respect to all the papers evaluated.

3.1 Ride Sharing Matching Algorithms

In a ride sharing system, it is expected that both the driver and the passenger can be joint into a shared trip. In other words, the drivers expect to find riders for their offered trips, and in turn, the passengers expect to find rides that supply their mobility necessities.

The ride sharing matching agencies discussed in Section 2.3.2 of this thesis manage the matching between the *ride offers* and the *ride requests*. For this purpose, a ride matching algorithm is used to find a subset of ride offers that may supply a given ride request. The subset is composed by feasible ride matchings that satisfy several spatial-temporal constraints involved in a ride sharing matching problem (FURUHATA et al., 2013).

In general, the spatial constraints are imposed by both the pick-up location, which requires a common location where both driver and passenger can meet and then start the trip, and the drop-off location, which requires that the passenger should be dropped at his intended destination. In turn, the temporal constraints refer to the pick-up and drop-off times, including in some cases, a certain tolerance for delays. Many matching agencies may also include some other constraints such as user reputation or some social aspects, such as genders, age, common friends, driver' and passenger' reviews (GUIDOTTI et al., 2015; BERLINGERIO et al., 2017).

A ride sharing system is effective when it is able to find feasible rides to their users under the many constraints involved. A largely adopted solution for reaching this objective consists in relaxing spatio-temporal constraints of ride offers and requests (STIGLIC et al., 2016; WANG et al., 2016). By relaxing the constraints, a ride sharing system can represent some particular user's flexibilities and take advantage of them to increase the possibility of ride matchings between drivers and passengers. However, it can bring more complexity to the ride matching algorithms.

In the next sections, we discuss some of these strategies to improve efficacy by adding flexibility to the constraints involved in a ride sharing matching, specifically we group them into the following groups: (a) Temporal flexibility (Section 3.1.1); (b) Spatial flexibility with detour (Section 3.1.2); (c) Spatial flexibility with slugging (Section 3.1.3); (d) Multi-

hop ride sharing (Section 3.1.4); (e) Social constraints (Section 3.1.5); (f) Activity-based flexibility (Section 3.1.6).

3.1.1 Temporal Flexibility

A common approach to increase the number of matchings is to consider that both the participant drivers and passengers have temporal flexibility: they can specify an earliest possible departure time and latest possible arrival time.

A recent work showed that even a small flexibility in terms of desired departure time or maximum detour time can significantly impact the expected matching rate, especially when the number of ride offers in the system is not large (STIGLIC et al., 2016). They found out that for a proper operation of the dynamic ride sharing systems, drivers and riders need to be flexible in terms of departure and arrival times (at least 10–15 min depending on origin and destination locations).

Using a similar approach, in (WANG; DESSOUKY; ORDONEZ, 2016) the authors study how the optimal routes change as a function of incentives for ride sharing, for example, inclusion of HOV (High Occupancy Vehicle) lanes, they modified existing pick-up and delivery problem with time windows to consider changes in passenger travel time and the cost of the travel due to vehicle load. In their approach, each driver participating in ride sharing provides his/her start location, end location, earliest departure time, and latest arrival time. A 0-1 integer programming model is formulated to solve the problem optimally. The authors proposed an algorithm called *Adjust Pickup Time Algorithm* to reduce the total cost and the customer ride time. They consider that each ride request provides its start location, end location, time windows for pickup and delivery, and the number of people who need to be served. Their results show that, as a participant in ride sharing becomes more flexible in time, the less she/he should pay for his/her trip. Moreover, their experiment results confirm the intuitive assumption that, with more ride offers, higher is the possibility of finding ride sharing matches.

3.1.2 Spatial flexibility with route detours

Analogously, other approaches explore the spatial flexibility of users in terms of ride detours or walking distance to catch the ride. Detour ride sharing considers a scenario where the driver accepts to make a detour from her original itinerary if this brings a satisfactory ride matching (GEISBERGER et al., 2010; CICI et al., 2014; STIGLIC et al., 2016).

Geisberger et. al in (GEISBERGER et al., 2010) propose an innovative detour route planning algorithm that efficiently minimize the ride detour of ride requests with arbitrary starting and destination points. They address the ride sharing problem in a scenario involving vehicle trips between cities. Using a public dataset of ride sharing offers from

Germany, the results of their experiments show an improvement of the matching rate of 20% compared to a baseline matching algorithm which do not consider route detours.

Cici et. al in (CICI et al., 2014) develop an efficient algorithm for matching users with similar mobility patterns, considering a range of constraints, including social distance and route detour. *The algorithm is based on heuristics used to solve the Capacitated Facility Location Problem with Unsplit Table Demand* (KORUPOLU; PLAXTON; RAJARAMAN, 2000). Experiment results using a dataset containing mobility data extracted from 3G call description records from the city of Madrid indicate a significant overlap in people's commute. Furthermore, in a scenario where rides can be shared only with neighbors with nearby home and work, even with a modest detour of 1 km they have observed a great potential reduction of 59% of the single-occupancy vehicle trips present in their dataset.

Furthermore, in (STIGLIC et al., 2016), besides the temporal flexibility, the authors also investigate the detour flexibility, representing the willingness of driver to make a detour in order to supply a ride request. They use a hierarchical optimization approach in which maximizes first the number of matches and subsequently maximize the system-wide vehicle miles savings for this maximum cardinality matching. They instantiate the optimization problem as an ILP, solving it by using CPLEX [1]. Their results show that participant flexibility is a very important in easing the matching process, specially when there are a low number of participants. They concluded that for a proper operation of ride sharing system, drivers and riders need to be flexible in terms of departure and arrival, but, most importantly, drivers need to be flexible in terms of the detour that they are willing to make.

Similar, in (WANG et al., 2016), the authors also study possible incentive to take detours to pick up additional passengers to qualify for High Occupancy Vehicles (HOV) lanes or discounted toll rates. Experiment results indicate that passengers are more encouraged to take detours (participate in ride sharing) to save ride time as time savings on HOV lanes increase.

3.1.3 Spatial Flexibility with Slugging

Orthogonally, some works investigated the possibility that the passenger could walk to a meeting point to join a ride (KELLEY, 2007; MINETT; PEARCE, 2011; MA; WOLFSON, 2013; STIGLIC et al., 2015). In literature, this approach is also known as *slugging*.

In (KELLEY, 2007), the author publish a theoretical study that exploits slugging in areas with HOV is proposed. Their study focus on the Casual carpooling which correspond to a kind of the ride sharing where the matching between the driver and passengers is not established in advance but coordinated on the spot. The author also addresses some shortcomings associated with casual carpooling such as personal safety, the "free-rider" problem, and the maximization of the number of passengers sharing a ride.

[1] www.ibm.com/analytics/data-science/prescriptive-analytics/cplex-optimizer

Minett and Pearce (MINETT; PEARCE, 2011) investigate the impact of slugging in gasoline consumption. For this purpose, in their analysis they computed and compared the energy consumption by: (a) single occupant vehicles; (b) carpool vehicles; (c) and a mix of buses and single occupant vehicles. They estimate that slugging in San Francisco can save from 1.7 to 3.5 million liters of gasoline per year, much of which comes from the indirect impact on the rest of the traffic.

In (MA; WOLFSON, 2013), the authors formally define the slugging problem and propose two heuristics for the matching problem aimed at saving total distance traveled by vehicles, namely: *Greedy-Benefit* and *Greedy-AVG-Benefit*. They consider vehicle capacity and travel time delay constraints. Both heuristics work in an iterative way, for each iteration, the *Greedy-Benefit* chooses the ride offer with individual maximum benefit, intuitively the *Greedy-AVG-Benefit*, choose the ones that in average have maximum average benefit. They provide proofs of their computational time complexity, proving the NP-completeness of the problem. They performed experiments using real taxi cabs trips in Shanghai, and reported a saving up to 59% in the total distance traveled by vehicles, whereas the optimal slugging plan achieves at most 70% savings.

Stiglic et. al in (STIGLIC et al., 2015) have shown that by considering the possibility for passengers to walk to specific meeting points in a ride sharing system can substantially improve a number of ride matching. It also allows a driver to be matched with multiple riders without increasing the number of stops the driver needs to make. They modeled their problem as a maximum weight bipartite matching problem and implemented an algorithm that optimally matches drivers and riders in large-scale ride sharing systems with meeting points. In their approach, riders may have to walk a short distance and may have to plan their time more carefully so as to ensure that they arrive on time at the meeting point where they are to be picked up (it is unlikely that drivers will be willing to wait for a rider at a pickup point for more than a minute or two). Moreover, the use of meeting points makes matches feasible predominantly because it allows a smaller detour for the driver (only in a few cases, it makes rider and driver time windows compatible). Without meeting points, approximately 90.6% of the riders have at least one feasible match. With meeting points, this fraction increases to approximately 92.5%. They also conclude that, as the number of meeting points increases, the number of feasible matches grows steadily.

3.1.4 Multi-hop Ride Sharing

Other works investigate how to combine more than one ride offer to supply a single ride request (HERBAWI; WEBER, 2012; DREWS; LUXEN, 2013; LIN et al., 2016; MASOUD; JAYAKRISHNAN, 2017). This approach is called multi-hop ride sharing. Basically, if a ride request can be matched with only one ride offer, then the problem is called single-hop

ride matching. It is called multi-hop ride matching, if a request can be matched with two or more ride offers at different times.

In (HERBAWI; WEBER, 2012), the authors have modeled the multi-hop ride matching problem with time windows and provided a genetic algorithm to solve it. From their experiment results, they conclude that the use of multi-hop ride matching can increase the number of ride matching with the penalty of increasing the total travel time for riders and drivers.

Drews and Luxen (2013) in (DREWS; LUXEN, 2013) presented an graph search algorithm that solves the time-dependent multi-hop ride sharing problem with a fixed set of stations. They modeled the multi-hop ride sharing considering that users travel between a number of stations that is related to timetable networks for public transportation. Additionally, they developed data structures and algorithms to efficiently compute matches. One interesting result of their algorithm is that increasing the number of transfers to more than two does not lead to significantly superior results on average anymore. However, the authors do not have specific contribution for the optimization of the matching problem, but instead they developed data structures and algorithms to efficiently compute matches.

In (LIN et al., 2016), the authors introduce multi-modal ride sharing considering the possibility of the passenger walk to a meeting point in a multi-hop (multiple drop-off) scenario. Their approach does not consider route detour and is mainly aimed for transportation hubs, such as airports, railway stations, etc. Their approach consists of two stages: (1) construction of a shareability graph, and (2) finding the maximum matching using such graph. The first stage finds all the possible pairs that can be merged in a way that satisfies the constraints of the two trips. In turn, the second phase, for an arbitrary graph, they search for the merging of pairs which results in the minimum number of merged trips in the pool. For finding the maximum matching, they use a standard existing algorithm (GALIL, 1986). They evaluated their approach by using 1.8 Million trips originated from La Guardia Airport in New York City. The experiment results indicate a trip-savings of about 25% when 75% of the passengers are willing to share the ride. Additionally, they found out how walking is valuable in combination with multiple drop-off ride sharing. For example, if passengers allow a 10-minutes walk, then the trips-reduction by ride sharing increases from about 10% to about 30%. Considering that at airports passengers often walk for 10 minutes from the gate to the curb, this assumption seems reasonable.

In (MASOUD; JAYAKRISHNAN, 2017), the authors discuss the features of a flexible ride sharing system and propose an algorithm to optimally solve the ride-matching problem in a flexible ride sharing system in real-time. They propose a multi-hop system with the ability to find itineraries for riders by means of optimally routing drivers. They implemented an optimal and real-time ride-matching algorithm using dynamic programming that maximizes the number of served riders in the system, while making the trips as

comfortable as possible by taking into consideration users' preferences on whom to ride with, and by minimizing the number of transfers and waiting times for riders. Furthermore, their results suggest that allowing transfers can have a considerable impact on the number of served riders.

3.1.5 Including Social Aspects in Ride Sharing

Recent works on carpooling and journey planning take into account, besides the spatial and temporal constraints, also social constraints (CICI et al., 2014; GUIDOTTI et al., 2015; CAMPANA; DELMASTRO; BRUNO, 2016; BERLINGERIO et al., 2017).

In (CICI et al., 2014), the authors focused their analysis also taking into account social constrains such as friendship in social media. Their motivation is that people often hesitant to ride with strangers. They inferred the social ties from call data records (CDRs), or declared friendship (Twitter). In the city of Madrid using CDR and Twitter direct friendship provides only a tiny traffic reduction of 1.1% and 1.2% respectively. However, when relaxing the social constraints and allowing ride sharing with friends-of-friends, the ride sharing potential increases significantly to 19% and 8.2% for friendship based on CDRs and Twitter data, respectively.

In (GUIDOTTI et al., 2015), the authors' methodology was tested on real data from Rome and San Francisco. The authors define a multi-objective optimization model that using a greedy approach first minimizes the number of cars, and afterwards tries to maximize the enjoyability of the user. For this purpose, the authors introduce a measure of enjoyability based on people's interests, social links, and tendency to connect to people with similar or dissimilar interests. They evaluate the approach in terms of cars saved, and average enjoyability of the users. From a study with more than 200 users reporting an interest of 39% in the enjoyable solution. Moreover, 24% of people declared that sharing the car with interesting people would be the primary motivation for carpooling.

The authors of (CAMPANA; DELMASTRO; BRUNO, 2016) address non-monetary aspects and social considerations that may influence the individual willingness of sharing a ride. They propose a recommender system for carpooling services that leverages on learning-to-rank techniques to automatically derive a personalized ranking model for each user from the history of her choices (i.e., the type of accepted or rejected shared rides). The system builds the list of recommended rides by maximizing the estimated success rate of the offered matches extracted from Foursquare check-in information. The results show that the proposed solution quickly obtains an accurate prediction of the personalized user's choice model.

In (BERLINGERIO et al., 2017), the authors extend the work (GUIDOTTI et al., 2015) by introducing a measure of enjoyability based on people's interests, social links, and tendency to connect to people with similar or dissimilar interests. Specifically, their enjoyability measure takes into account two factors: (i) like-mindness, i.e. a topic similarity

between any two users; and (ii) homophily, i.e. the tendency of a person to group with similar ones. The authors address a ride sharing problem where they try to minimize the number of recurring trips made by cars and also maximize the enjoyability the trip considering social aspects. In their experiments, they do the extraction of enjoyability and mobility demand from Twitter. Their approach save up to 57% of the cars, while the total enjoyability is up to double.

3.1.6 Activity-Based Ride Sharing

Most of related works do not consider in their models the activity as a flexibility attribute, except the work by Cho et Al. (CHO et al., 2013) that first addresses activity-based carpooling. The authors propose the use of an ontology in an activity-based microsimulation. While no explicit evidence is presented, the focus of the paper is recognizing that the ontology is a useful and appropriate method for activity-based microsimulation research. Indeed, only a conceptual design and framework are suggested, and this study is a clearly preliminary step.

In (WANG; KUTADINATA; WINTER, 2016), they propose an algorithm that expands the potential destination choice set by considering alternative destinations that provide a similar activity function as the originals. The matching is conducted considering a static pre-planed daily schedules of all participants involved. The authors define a daily schedule as a composition of multiple trip chains. Given the combinatorial computational complexity in deciding the destination choice set for a chain of multiple flexible activities, the author introduce a space-time filter algorithm to search for feasible rides. The algorithm considers a reasonable time window, while still allowing a detour tolerance for each trip. A global optimal matching is achieved by binary linear programming. The experiments confirm the capability of activity-based ride sharing to increase successful matching rates.

3.1.7 Comparative Analysis

Table 1 summarizes the works previously discussed indicating the approaches used to represent participants' flexibility (e.g. temporal, spatial with detour, spatial with slugging, multi-hop, social and activity-based) and their objectives. We can observe that most of the works in literature relax the temporal constraints by representing the time as window intervals. Moreover, the spatial flexibility plays an important role on finding matches between drivers and passengers. This flexibility can be represented either by allowing the drivers to perform detour on their trips or by considering that the passenger can walk to a specific meeting point. The multi-hop ride sharing can increase the amount of possible matchings, however it may cause more discomfort to the passenger given the necessity of switching between vehicles.

Similar to the works presented above, we try to increase the matching possibilities by relaxing the constraints involved in ride sharing scenario. Particularly, we impose a higher flexibility to the spatial constraints, by exploring a new dimension based on the intended activity of the passenger. This represents the possibility of a passenger to choose one among multiple destination choices where to go to perform her intended activity. Thus, the approach presented in this thesis is completely orthogonal and possibly complementary to the studies presented in this Section. For example, both approaches activity based and route detour could be used by matching agencies to represent respectively the flexibility of the passenger to have multiple destination options, and the flexibility of the driver to perform detour on her original route to supply a ride request.

This thesis has similar motivations to Wang's (WANG; KUTADINATA; WINTER, 2016). In (WANG; KUTADINATA; WINTER, 2016), the authors also propose an approach for expanding the potential destination choices set by considering alternative destinations looking for places that provide similar activity function as the original. Among the difference with the current work we can point out:

(a) They do not propose any mechanism to evaluate the quality of the ride in fitting the requirement of the participants. We propose a ranking model that compute a set of features from the ride options, scoring with high values the ones that best matches the ride request requirements; (b) We use a slugging ride sharing approach where we consider that the passenger is able to walk until the pickup-up point, while they use a detour considering that the driver can make a detour on his route. In our study, we use slugging to support a sustainable idea that people transportation can benefit from daily route of other people; (c) Finally, they focus on finding a maximal optimal combination of matches for a day schedule, assuming to know in advance all the daily schedule of the participants. Instead, we do not consider the whole day trip of the users. We focus on a more realistic scenario to optimize a set of on demand ride request rather than daily schedule of users.

Table 1 – Literature on ride sharing matching algorithm

Work	**Temporal**	**Detour**	**Slug.**	**M.Hop.**	**Social**	**Activity**	**Objective**
(STIGLIC et al., 2016)	yes	yes	no	no	no	no	Max. number of matches Max. driving kms savings
(WANG et al., 2016)	yes	no	no	no	no	no	Total cost and ride time
(GEISBERGER et al., 2010)	yes	yes	no	no	no	no	Min. route detours
(CICI et al., 2014)	yes	yes	no	no	yes	no	Max. number of matches
(KELLEY, 2007)	yes	no	yes	no	no	no	Max. number of matches
(MINETT; PEARCE, 2011)	yes	no	yes	no	no	no	Reduce energy consumption
(MA; WOLFSON, 2013)	yes	no	yes	no	no	no	Max. driving kms savings
(STIGLIC et al., 2015)	yes	no	yes	no	no	no	Max. driving kms savings
(HERBAWI; WEBER, 2012)	yes	no	no	yes	no	no	Max. number of matches
(DREWS; LUXEN, 2013)	yes	no	no	yes	no	no	Efficient matching computation
(LIN et al., 2016)	yes	no	no	yes	no	no	Max. number of matches
(MASOUD; JAYAKRISHNAN, 2017)	yes	no	yes	no	no	no	Max. number of matches
(GUIDOTTI et al., 2015)	yes	no	no	no	yes	no	Max. number of matches Max. enjoyability
(CAMPANA; DELMASTRO; BRUNO, 2016)	yes	no	no	no	yes	no	Max. success rate of the offered matches
(BERLINGERIO et al., 2017)	yes	no	no	no	yes	no	Max. number of matches Max. enjoyability
(WANG et al., 2016)	yes	yes	no	no	no	yes	Max. number of matches
Current Work (LIRA et al., 2015; LIRA et al., 2018)	yes	no	yes	no	no	yes	Max. number of matches Max. satisfaction

3.2 Exploring Event Attendance Using Social Media

Large events like music festivals or religious celebrations attract thousands of participants. In general, such events raise concerns regarding the mobility of the attendees to the area of the event. A common practice is to provide transportation means to encourage people to switch their mode of transportation from solo to shared driving. For example, if most people arrive at the event destination via private cars, this creates traffic congestion and parking overloading in areas around the event. As a means of overcoming this problem, ride sharing transportation service operators (Section 2.3.1) as shuttle bus perform circulated routes to supply the mobility demands to attendees of the event. From the logistic perspective, the problem comprises the identification of the demand of rides to proper allocate the transportation vehicles. However, especially for large events, it is not a simple and straightforward task to identify such demands (FEEHAN, 2006).

Chasing this problem, this thesis introduces an approach to infer such mobility demand by investigating users' attendance through the use of social media. Recent researches have shown that social media play a role in understanding modern life, including transport (EFTHYMIOU; ANTONIOU, 2012; GAL-TZUR et al., 2014b; D'ANDREA et al., 2015) and human mobility analysis (CESARIO et al., 2016; MIKUSZ et al., 2016; HAWELKA et al., 2014). Large events are usually well reflected in social media where interested users express, through *posts*, their feelings, experiences or opinions about such events. The social media network can therefore be seen as a channel between people and "the event" through the posts referring to the event using an *hashtag*, or the event's name or its user handle. Therefore, we want to explore social media as a source of potential user event attendees to understand mobility demand.

For this purpose, we rely on the post content to infer user' attendance to large events. In literature, many papers tackle the problem of estimating the current location of users or their home from non geo-located posts (CHENG; CAVERLEE; LEE, 2010; CHANG et al., 2012; MAHMUD; NICHOLS; DREWS, 2014; LEE et al., 2014; KINSELLA; MURDOCK; O'HARE, 2011; ONAN, 2017; BAKERMAN et al., 2018; EFSTATHIADES et al., 2016). Compared to these proposals, we have a different objective as we do not want to estimate the exact user location at the time of the post, but classify the single posts on the basis of user future, current and past attendance to a given event. Understanding demand is an initial and important phase for transportation demand management (ISLAM et al., 2016). Specially for large events, understanding the mobility demand may lead to more accurate transportation services.

Therefore, we compare our approach with existing works in literature that study and analyze event attendance using social media. Events in social media have been extensively studied. The main aspects investigated in the literature are: (1) prediction of events attendance in Event-Based Social Networks (EBSN) and Location-Based Social Network (LBSN) (DU et al., 2016; ZHANG; ZHAO; CAO, 2015; GEORGIEV; NOULAS; MASCOLO, 2014);

(2) recommendation of events to users (QUERCIA et al., 2010; MACEDO; MARINHO; SANTOS, 2015; GAO et al., 2016); (3), estimation of the number of attendees in a given event (BOTTA; MOAT; PREIS, 2015), and, (4) modeling participants' behavior during an event (CESARIO et al., 2017; CESARIO et al., 2016; CESARIO et al., 2015).

3.2.1 Prediction of Events Attendance in EBSN and LBSN

Du et al. (DU et al., 2016) analysed an EBSN to predict users' attendance by taking into account the content, the spatial and temporal context, the users' preferences and their social influence. They used a *singular value decomposition with multi-factor neighborhood (SVD-MFN)* algorithm to predict activity attendance on the Douban Events network. Zhang et al. (ZHANG; ZHAO; CAO, 2015) proposed a supervised learning model to predict event attendance based on semantic, temporal, and spatial features, representing how frequently and when users attended similar events in the past, the semantic similarity between events, the location preference when attending events and the home location of the user. They trained three classifiers on a Meetup dataset with semantic descriptions of all events organized.

Georgiev et al. (GEORGIEV; NOULAS; MASCOLO, 2014) addressed the extent to which geospatial, temporal, and social factors influence the users' preferences towards events formulating a predictive modeling task trying to match a user's mobility profile against the collective past Foursquare check-in activity of potential event attendees. Zhang and Lv (ZHANG; LV, 2017) proposed a group-based social influence propagation network to model group-specific influences on events. In (ZHANG; LV, 2018), the same authors extended the previous work proposing a group-based event participation prediction framework that embeds and connects group context features and social related features using historical event attendance logs. The authors extract the group-based social features by using a hybrid event-group/category-user network that captures intrinsic social relationships. Their results show that these features are important for predicting event participation.

Compared to these approaches, we do not specifically deal with EBSN and LBSN, but instead we focus on popular social media where events can have an "echo". We do not use users history or preferences as we aim at classifying single posts by disregarding the user profile and specific events information.

3.2.2 Recommendation of events to users

Within the second category, event recommendation, papers (QUERCIA et al., 2010; MACEDO; MARINHO; SANTOS, 2015; GAO et al., 2016; MO et al., 2018) and (WANG et al., 2016) addressed the challenge of recommending events within event-based social networks (EBSNs). Each of these approaches is challenged by the cold-start problem, and recommendation evidence may resort to the events that are geographically closest (QUERCIA et al.,

2010). The works in (GAO et al., 2016) and (LIU et al., 2017) study the influence of social groups to improve the event recommendation performance. Gao et al. (GAO et al., 2016) propose a new Bayesian latent factor model that combines social group influence and individual preference for event recommendation. In turn, Liu et al. (LIU et al., 2017) propose a collective pairwise matrix factorization model to estimate users' pairwise preferences on events, groups and locations.

Macedo et al (MACEDO; MARINHO; SANTOS, 2015) propose a recommendation approach that leverages multiple context-aware recommendation models for learning to rank events. They exploit further features based on group memberships, location signals based on the users' geographical preferences, and temporal signals derived from the users' time preferences. In (MO et al., 2018), the authors consider also the capacity of an event to limit the number of users for the recommendation. Their objectives is to coordinate unbalanced user arrangements among the recommended events. The works in (WANG et al., 2016) and (QIN; RISHABH; CARNAHAN, 2016) focus on efficient and scalable learning technique for event recommendation to handle large-scale, straming data. Our work is complementary with respect to these approaches since we are interested in identifying the posts related to event attendance rather than in making recommendations. In any case our approach could allow to identify more precisely the target users for recommendations.

3.2.3 Estimation of the number of attendees in a given event

Within the third category of related works, Botta et al. in (BOTTA; MOAT; PREIS, 2015) investigated whether mobile phone usage and the geolocated Twitter data can be used to estimate the number of people in a specific area at a given time. They consider two case studies of access-restricted areas in Italy: a stadium and an airport (where there were ground truth visitor statistics), they concluded that geolocated tweets with mobile phone data could be a good proxy of estimating the number of users. Sinnott and Wang provide solutions to estimate the population of suburbs and skyscrapers through the use of geo-tagged Twitter data (SINNOTT; WANG, 2017). They construct linear models for suburbs of four cities and investigate spatial correlation properties between the geo-tagged tweets and the official Census data. Their results show that Twitter can be used for micro-population estimation with quantifiable degrees of accuracy.

In (SINNOTT; CHEN, 2016b), the authors propose a regression model to estimate the number of attendees from the amount of geo-tagged tweets posted at an event. They apply the prediction model to estimate the attendance at the Melbourne marathon.

3.2.4 Modeling participants' behavior during an event

Finally, in the last category of works, the authors of (CESARIO et al., 2016; CESARIO et al., 2015; CESARIO et al., 2017) described a methodology for identifying the user behavior and

mobility patterns of Instagram social network users visiting the EXPO 2015 world fair in Milan and the FIFA World Cup 2014. They analysed how the number of visitors changes over time, identify the most frequent sets of visited pavilions, which countries the visitors came from, and the main destinations of foreign visitors to Italian regions and cities after their visit to the EXPO. They also analysed geotagged tweets of people attending the 2014 FIFA World Cup identifying the most frequent movements of fans, the number of matches attended by groups of fans, clusters of most attended matches, and the most frequented stadiums.

3.2.5 Comparative Analysis

These latter two groups of works have similar objectives to our aim in studying the social media users' actual participation in events. However, the main differences are: (1) we do not use geotagged information to identify current attendance, but we rely on the media posts content to infer users' participation in events. Compared to the related works based on geotagged data, we explore a higher number of posts about the event since a low percentage of the social media posts are geotagged. For example, on Twitter, around 2% of the tweets are geotagged[2]. Moreover, our approach allows the analysis of event attendance also based on post made before, during and after the event; (2) we do not estimate a global number of participants or crowd, but instead we identify specific social media users who are likely to be – or have been – present at an event. Additionally, our final objective is to infer user attendance to derive the mobility demand to large event rather than derive event crowd size. Our approach can thus provide useful and complementary information to support both applications crowd behavior modeling and crowd size estimation in large events; and (3) we do not recommend participation but instead we infer current, future or past attendance of users based on the media posts.

3.3 Conclusions

In this chapter, we have reviewed several works that are relevant and have common purposes with the objectives of this thesis:

- We have evaluated some **ride sharing matching algorithms**.These approaches have in common the aim of increasing the number of ride matches, focusing on maximizing either the number of participants or the driving distance savings. Recent approaches consider also social aspects of the participants to improve enjoyability of the ride trips. However, most of the approaches in literature consider only the spatial-temporal aspects when matching ride offers and ride requests, ignoring the semantic behind the mobility demand necessities. Thus, orthogonally to the current

[2] http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654

approaches, this thesis present a novel approach for ride matching based on the intended activity of the passenger. This represents a new way for relaxing the spatial constraints, exploring alternative destinations to find feasible ride sharing matchings. The alternative destinations are proposed based on the indented activity of the passenger. Moreover, given the orthogonality of this proposed approach, ride sharing matching agencies could apply it jointly with other matching strategies as route detour, slugging and multi-hop ride sharing.

- Finally, we have discussed several papers that investigate **event attendance through the use of social media**. We have discussed the main groups of works in this branch of research. The most similar works rely on geo-tagged information to infer event attendance. These approaches have two limitations: (1) a small percentage of posts in social media have geo-tageed information; (2) by considering only posts made during the event, but not the ones made before, future attendance prediction analysis is not enabled. Thus, our proposed approach to infer event attendance overcomes these limitations providing a novel way to understand mobility demand through the use of social media.

4 The Activity-Based Ride Matching (ABRM) Algorithm

Recent studies of human mobility highlight an individual characteristic of the people: the tendency to be regular or not in choosing the places where to perform some activities (LIRA et al., 2014; WU; LI, 2016). Be regular here refers to the fact that the person often choose the same place to perform a given activity. While not be regular represents a not uniform behavior of a person on choosing a place for doing a given activity. Indeed, we notice that people can perform their activities at different alternative locations. For example, think about the different shopping malls where to go shopping, or the different Italian restaurants where to go for dinner.

In a ride sharing system essentially there is a matching phase between drivers and riders. In general, it means that the paths performed by these people have moving stretches of common interest, and therefore they can share a vehicle for a given trip. In most of the cases the destination is fixed and cannot be changed (e.g. go to work or go home). Here we propose a different scenario where the destination is not fixed and can be changed if the activity to be done (e.g. shopping, eating ...) is preserved.

We propose the Activity-Based Ride Matching (ABRM) algorithm: we focus on the user's desired activity, rather than only considering the path or final destination of the rides thus increasing the number of possible matches. In order to present the proposed matching algorithm, this chapter is organized as follows: Section 4.1 introduces some basic definitions used in this chapter and defines our Activity-Based Ride Matching Retrieval Problem. Section 4.2 describes the algorithm for matching ride requests with alternative destinations and discusses the ranking model adopted. We discuss the experiments in Section 4.4. We also introduce in Section 4.5 a prototype called ComeWithMe that implements the Activity-Based Ride Matching Algorithm in almost its entire query pipeline. Finally, we draw the final considerations on the current chapter in Section 4.6.

4.1 Basic Definitions and Problem Formalization

We introduce here the formulation of our Activity-Based Ride Matching *Retrieval* problem and some basic concepts and notations used through the thesis.

A *trajectory* (or *trip*) represents the spatio-temporal movement of a traveling object. It is generally recorded by a tracking device into samples thus bringing a discrete representation of the movement. A more formal definition is given below.

Definition 1 (trajectory)**.** *A trajectory represents a finite sequence of spatio-temporal*

points assigned to a moving object and denoted by $<objId, < x_1, y_1, t_1 >, \ldots < x_n, y_n, t_n >>$, *where* $objId$ *is the identifier of the moving object, and* x_i, y_i, t_i *represent the spatial and temporal coordinates of the sample points.*

Trajectories may have stops at venues or Points Of Interest (POIs), e.g., a shop, a bar, a restaurant or a gym (PARENT et al., 2013):

Definition 2 (Point Of Interest (POI) or venue)**.** *A POI is a geographical object, usually associated to a human activity, that is interesting for a specific application. We define a POI as a tuple* $< s, n, c >$ *where* s *is the representative spatial point,* n *is the name of the POI and* c *is its category drawn from a defined taxonomy* C*.*

In the real world, some POIs may have more than one activity associated to them. For example, in a shopping mall, it is usually possible to perform activities like shopping, eating and ATM services. In this thesis, for the sake of simplicity, we consider that a POI is related to only one activity. The chosen activity is the one associated to the primary category of the POI. Thus, considering our previous example, for a shopping mall, the associated activity is shopping.

A trajectory can pass by or stop close to a set of POIs where some *activity* (e.g., shopping, visiting, eating, working, exercising) can be performed.

Definition 3 (Activity)**.** *An Activity defines a task that can be performed at a POI. We assume that activities are related to specific POI categories and viceversa. Thus there is a mapping between a given activity a and a set of POI categories. For example, the activity eating is related to POI categories Restaurants and Pizzerie. Viceversa the POI category Restaurants is related to activities eating and drinking.*

Let P be the set of all POIs. Given a POI $p \in P$, we thus assume to be able to find a number of alternative venues that is a small subset of P where the activity performed in p can be performed as well. We call these POIs the *alternative destinations* for p.

Definition 4 (Alternative Destination)**.** *Given a POI* p*, the set of alternative destinations for* p *is a set of POIs* $< p_1, \ldots, p_n > | \ \forall p_i \in P$ *where it is possible to perform the same activity as in* p*. These alternative destinations can be all the POIs belonging to the same category of* p *or a subset of them selected on the basis of some* criterium*, e.g., those most similar to* p*, or the most popular, or the ones preferred by the user.*

It is important to note that the set of alternative destinations for a given POI p can be an empty set. This implies that a person visits the POI p to perform a very specific activity, thus not being flexible to consider other alternative destinations. For example, the work place of a person is in general unique, not being possible to look for alternative destination for this category of place. The same restriction can be considered for POI categories like airports, gyms, universities, etc.

The ABRM problem assumes that a *ride request* for going to a POI p from a given spatial location *loc* at a given time t can be served with *ride offers* respecting the time and starting location constraints of the request and possibly dropping the user to one of the alternative destinations for p. Thus, we formally define the terms *ride request* and *ride offer*.

Definition 5 (Ride Request)**.** *We represent a Ride Request q as a tuple: $< u, loc, p, time, w\_dist, delay >$, where u identifies the requesting user, $p \in P$ is the POI to be reached, loc and time are the starting location and the preferred departure time, while w_dist is the maximum walking distance the user is willing to walk to get the ride and delay the maximum time the user is willing to anticipate or delay the departure.*

Definition 6 (Ride Offer)**.** *A ride offer rt is a tuple: $< u, orig, dest, time, path >$ where u is the driver, orig and dest are the fixed origin and destination of the ride offered by u, time is the departure time, and path is the route followed by the vehicle offering available seats. These seats can be offered to passengers traveling from origins to destinations that are reachable along the vehicle route.*

It is worth noticing that our definition of ride offer is intentionally generic to encompass different ride sharing scenarios: the vehicle with available seats following a fixed route at a fixed time could be a private car (e.g., routinely going home from work at 5pm of working days), or a shared taxi serving a ride.

Thus, we can now proceed in formulating our Activity-Based Ride Matching problem:

Definition 7 (Activity-Based Ride Matching Problem)**.** *Given a set of ride offers RT and a ride request $q =< u, loc, p, time, w\_dist, delay >$, the ABRM problem seeks to find all the **ride matchings** $\mathcal{M} : \{m_1, \ldots m_k\}$ between q and the ride offers in RT that allow the passenger to reach p or an alternative destination for p within the maximum walking distance w_dist and the maximum departure delay or anticipation delay. Thus a ride matching $m \in \mathcal{M}$ satisfies the following constraints:*

1. The walking distance for the passenger to reach the pick-up location cannot be higher than the maximum walking distance specified by in the ride request q.

$$distance(q.loc, m.pickupLoc) \leq q.w_{dist} \tag{4.1}$$

2. Analogous, the walk distance for the passenger to reach the POI destination $p \in A^q$ cannot be higher than the maximum walking distance specified in the ride request.

$$distance(m.dropLoc, p.loc) \leq q.w_{dist} \quad (p \in P^q) \tag{4.2}$$

3. The pick-up time may be delayed or anticipated respecting a maximum limit in minutes $q.delay$.

$$|q.time - m.pickuptime| \leq q.delay \tag{4.3}$$

4. The pickup time must be before the drop-off time.

$$m.pickuptime \leq m.droptime \tag{4.4}$$

4.2 Activity-Based RideMatching Algorithm

The example in Figure 8 illustrates a simple instance of our matching problem. The request q of user u is for POI v_1 starting from location p at time t with a maximum walking distance of $500mt$ and a temporal flexibility of 30 min. The alternative destinations for v_1, preserving the activity to be done, are POIs v_2, v_3, v_4, v_5. The circles around the POIs represent the area within the walking distance the user set in her request (e.g., 500 meters). The set RT does not offer any ride to v_1, v_2 and v_3 satisfying the time and starting location constraints of u. There is, however, a ride offer $rt \in RT$, that starts at time t_o and ends at time t_f, intersects the circles around p, v_4, or v_5 in the order and respects the temporal constraint $t_o < t < t_f$. Notice that rt represents a trajectory with its orientation according to the arrow.

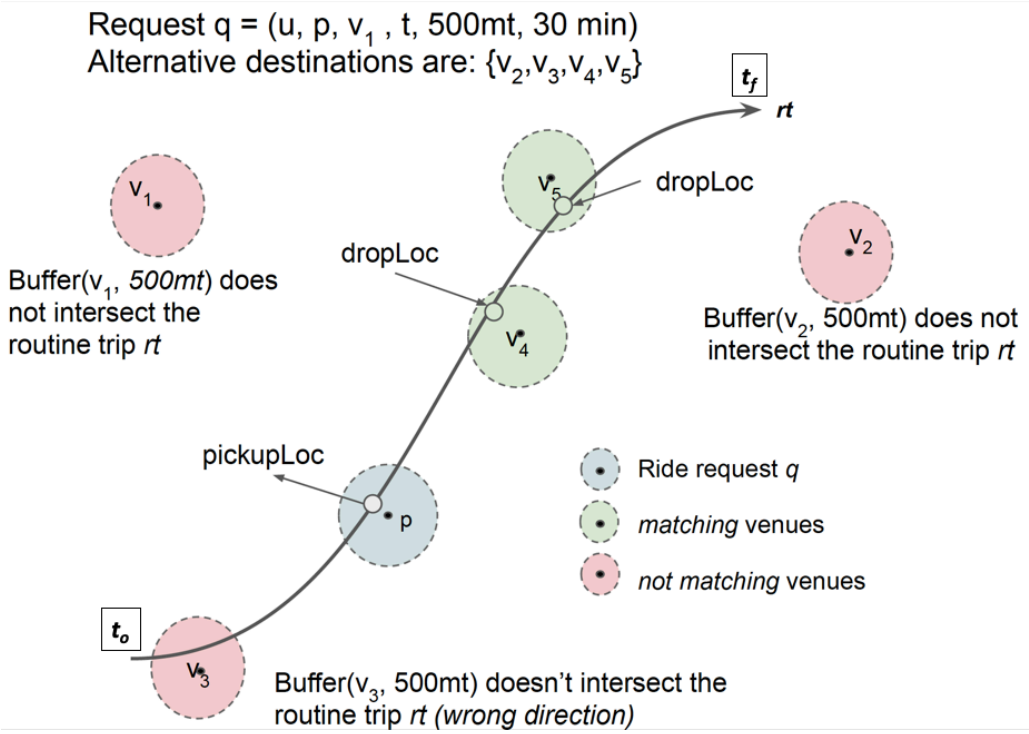


Figure 8 – A *ride matching* example

Algoritmo 1: Activity-Based Ride Matching

```
   Input  : q = (u, loc, venueDest, time, w_dist, delay) % ride request
            RT % set of ride offers
            V % set of POIs
   Output: M % set of matchings
 1 begin
 2     M ← ∅;
 3     altVenues ← AlternativeDest(q.venueDest, V);
 4     foreach rt ∈ RT do
 5         if Distance(q.loc, rt.traj) ≤ q.w_dist then
 6             m.pickupLoc = closestPoint(q.loc, rt.traj);
 7             m.pickupTime = timeAt(rt.traj, m.pickupLoc);
 8             if (|q.time − m.pickupTime| ≤ q.delay) then
 9                 foreach vdest ∈ altVenues do
10                     if Distance(vdest.s, rt.traj) ≤ q.w_dist then
11                         m.dropLoc = closestPoint(vdest.s, rt.traj);
12                         m.droptime = timeAt(rt.traj, m.dropLoc);
13                     if m.droptime > m.pickupTime then
14                         m.dest = vdest;
15                         M ← M + m;
16     return M
```

We have thus two possible matches: one possibility is to pick up the ride rt from *pickupLoc* to v_4 and another possibility is take the same ride up to v_5. We call these ride possibilities the *matchings*. We note that a request may have different matchings, not only with different ride offers but also with the same ride to different alternative destinations. Each matching has a destination venue *altDestination*, a pickup location *pickupLoc* and time *pickupTime*, a drop-off location *dropLoc* and time *dropTime*. We specify that our approach is not based on the detour of the ride, that remain fixed, and the driver does not change her path to pickup the passenger. It is the passenger who moves from the location of the request to the *pickupLoc* point of the matching. The constraint is that *pickupLoc* and *dropLoc* have distances from *loc* and *altDestination* lower than the walk distance *w_dist* specified in q.

The pseudocode in Algorithm 1 illustrates the steps that find all the matchings $\mathcal{M}$ for a user request $q = (u, loc, venueDest, time, w\_dist, delay)$ in a set RT of ride offers.

The algorithm starts by finding the alternative destinations for *venueDest* with function *AlternativeDest()*. This function can be instantiated in several ways. A simple solution is to select all the POIs whose category is the same as the requested POI. For example, when the user requests a ride to a supermarket, the alternative POIs are all the venues labeled with the "supermarket" category. Since the POIs in a given category can be many, we can restrict the number of alternative venues by choosing the k most *popular* ones or the k most similar to *venueDest* according to some similarity function.

The algorithm then iterates over the ride offers (line 4) checking if the starting location *q.loc* is within the walking distance from the path of a ride offer (line 5). In this case, the algorithm sets as pick up location the closest point between the offered ride path and the request location (line 6). Accordingly, the pick up time is computed as the time at which

the offered trip passes in the pick up location (line 7). Then, if the candidate ride does not respect the time delay constraint for pick up it is discarded (line 8).

A similar computation is done for the drop point. The algorithm checks if the candidate ride passes within distance *w_dist* from one of the alternative destinations (line 9-12) and checks that the direction of the ride is correct (line 13). In this case, the candidate ride is added to the matchings set (line 15). Finally, the algorithm returns the set of matchings found (line 16).

Since the cardinality of the set of ride matching can be high, we introduce a subsequent ranking step to order the matchings on the basis of an estimation of their relevance for the passenger (or on the basis of other criteria such as the saving of CO_2). In the next section we discuss the ranking features and how we can combine them in a ranking function.

4.2.1 Ranking Model

In order to rank the matchings returned by Algorithm 1 we consider four features. The objective of these features is to measure how effectively a matching m fits a ride request q. The four features are:

Time delay (f_{dly}). The *anticipation* or *delay* of the trip respect to the intended time specified in the request.

Distance to walk (f_{wld}). The *distance* the passenger has to walk in order to get the ride and arrive at the proposed destination. It is computed as the sum of the distance between the passenger location and the pick-up point, plus the distance between the drop off point and the destination venue.

Ride duration (f_{dur}). The estimated *duration* of the ride from the user location to the proposed destination.

Ride length (f_{len}). The estimated *length* of the candidate ride.

These features are first rescaled in the range [0-1] on a ride-request basis by considering all the values occurring in the set of matchings $\mathcal{M}$. Thus, the highest value of the feature has value equals to 1, while the lowest is equals to zero. $\mathcal{M}$ is then sorted by decreasing value of function $Rank(m)$:

$$Rank(m) = 1 - \sum_i w_i f_i$$

where $w_i \in [0, 1], \sum_i w_i = 1$ is the weight associated with feature $f_i \in \{f_{dly}, f_{wld}, f_{dur}, f_{len}\}$. By properly setting the weights w_i in the above linear combination of features, we can tune the importance of the different features. For example, we can reward rides with the shortest distance to reduce the emission of pollutants, or favor the passengers by ranking higher rides with the lowest walks and/or duration. Of course we are aware that more sophisticated ranking models considering different and complex aspects could be adopted (see for example (CAMPANA; DELMASTRO; BRUNO, 2016)). However, the goal of this work is investigating the impact on carpooling of user flexibility and not studying application-

specific rankers. We will see in the following how this simple ranking model allows us to sweep weights in order to understand their effect on the user and the environment.

4.3 Research Questions

The main objective of this work is to increase the efficiency of the ride-sharing systems by exploring alternative destinations to supply ride requests. The reallocation to an alternative destination is based on the intended activity of the passenger. We investigate the impact of our proposed approach by four research questions. The experiments discussed at 4.4 aim to answer comprehensively the following research questions:

***R**Q1: To what extent can ABRM increase ride sharing opportunities?*
***R**Q2: How well the matchings discovered by ABRM fit the constraints in the ride request?*
***R**Q3: What is the impact of tuning the weights used by the ranking function on the number of requests potentially supplied?*
***R**Q4: Which are the most favorable activities for exploiting the alternative destinations approach?*

4.4 Experimental Evaluation

In this section we present the experiments conducted to assess ABRM in terms of kilometers, liters of gasoline and CO_2 potentially saved with respect to a traditional, destination-oriented, carpooling approach. Note that in the following we do not deal with the allocation problem neither consider the number of seats available in the cars offering the rides. Ride request-offer allocation is a well-known optimization problem, orthogonal to this proposal. Any scheduling solution addressing this problem for destination-oriented carpooling fits also our activity-based approach. Since the same assumptions hold for the destination-oriented solution used as baseline, we believe that the choice of not considering allocation does not constitute a limitation of the work. Thus, below we investigate the potential impact of ABRM and of the settings of the ranking function in the reduction of the number of circulating vehicles, of pollutant emissions and consequent improvement of quality of the urban environment. The results of this work have been published in (LIRA et al., 2015; LIRA et al., 2018).

4.4.1 Experimental setup

Our experiments are conducted on two semi-synthetic[1] datasets of ride requests and ride offers, obtained by processing and enriching two publicly-available *Foursquare* datasets

[1] Here, the term "semi-synthetic" refers to the fact that we are using real check-ins performed by real users, but the trajectories between these checkins, the ride offers and the ride requests have been created by using a heuristic.

Table 2 – Datasets Statistics

Dataset	**Checkins**	**Users**	**Venues**	**Categories**
New York (NYC)	227,428	1,083	38,333	251
Tokyo (TKY)	573,703	2,293	61,858	247

(YANG et al., 2015). These datasets record the check-ins of *Foursquare* users in New York City (NYC) and Tokyo (TKY) for about 10 months (from 12 April 2012 to 16 February 2013). Each check-in is associated with a time stamp, the GPS coordinates of the POI and a fine-grained venue-category. Table 6 summarizes the total number of check-ins, the number of distinct users and distinct venues, and the number of possible categories for venues. For both datasets we consider only the users with at least 100 check-ins.

We enriched the above datasets by gathering the POI information using the *Venue*[2] and *Similar*[3] Foursquare APIs.Specifically, we gathered for each POI in our datasets the number of check-ins performed at the POI (popularity), the number of "likes" received (favorite), and the $top-5$ most similar venues according to an unknown Foursquare similarity measure.

We emphasize that the datasets used for the experiments provide only a simulation of a traffic scenario and are not representative of a general mobility graph. In these datasets the urban traffic flow is surely under-represented as most actual trips are likely not to be between two FourSquare destinations checked-in by the drivers. Nevertheless, this dataset has the important advantage that the activity performed by the users is explicitly reported as check-ins and this is crucial for our activity-based ride sharing approach. Rather than representing a general urban traffic flow, we simulate using these datasets the activity-based ride requests and offers. In the previous preliminary work (LIRA et al., 2015), we experimented the use of actual GPS traces of cars and we faced the non trivial problem of associating the raw GPS points to the performed activity at stops. Also we had strong privacy problems and we could not make the dataset public. Therefore we privileged here the use of datasets that, although semi-synthetic and with clear limitations, are public and representative of a large-scale activity-based scenario.

We exploit the above datasets of Foursquare check-ins to build two semi-synthetic datasets representing disjoint sets of *Ride Requests* and *Ride Offers*. By matching the ride requests with the ride offers by means of the ABRM algorithm and the baseline, we assess our proposal by considering the set of requests potentially satisfied by ABRM but not by the baseline algorithm.

We identify the ride offers as the trips of each user between the two most frequently visited venues v_a and v_b. The intuition behind this choice is that the rides between the most frequently visited locations constitute a reasonable surrogate of routine trips a user

[2] https://developer.foursquare.com/docs/venues/venues
[3] https://developer.foursquare.com/docs/venues/similar

could offer as driver. This is also supported by a manual inspection of the data that shows that very often people have their most frequent check-ins at venues such as home, work place, university, school (LIRA et al., 2014).

The exact procedure followed for each user u to populate the set RT of ride offers is detailed below:

1. Let v_a and v_b be the two POIs most frequently visited by user u, and rt_{ab} and rt_{ba} the candidate routine trips from v_a to v_b and from v_b to v_a, respectively.

2. rt_{ab} and rt_{ba} are added to set RT for all the days of the week in which there is at least a check-in of u in both the places. The arrival time in v_a and v_b for the above two ride offers are computed as the median among the timestamps associated with the check-ins in v_a and v_b.

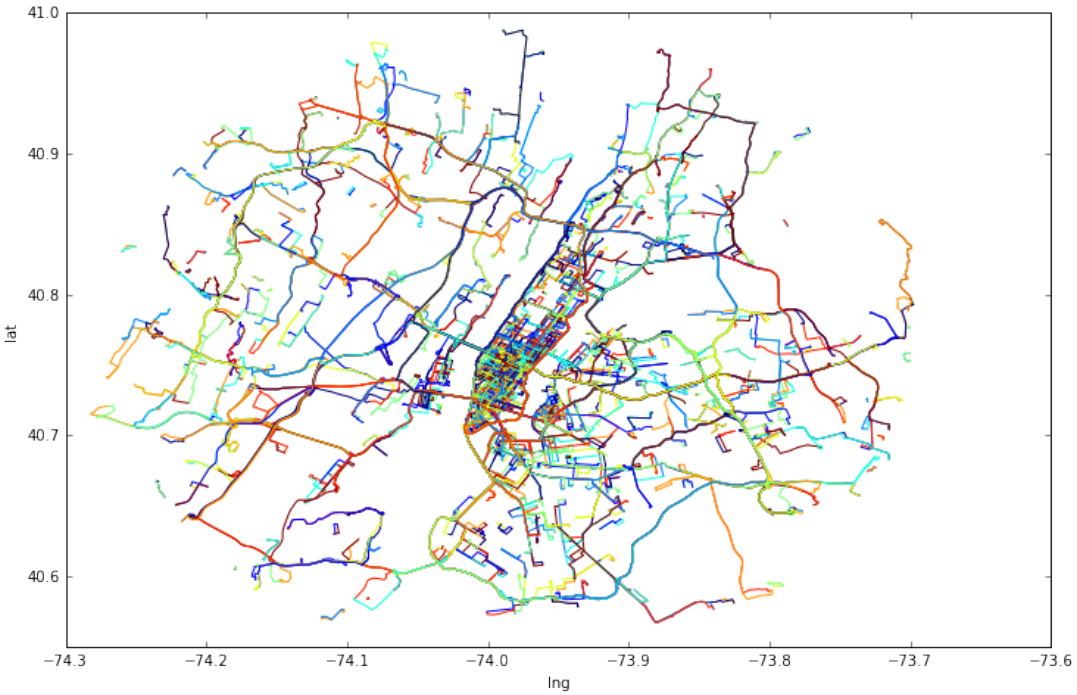
3. For each ride offer rt obtained with steps 1-2 we compute a representative trajectory of the fastest car route from the departure to the arrival locations by using *Google Maps*[4]. In addition, the arrival time $t_{arrival}$ of rt is used to estimate the duration of the ride, its length and the time of departure $t_{departure}$.

By following the above procedure we obtained 11,426 and 25,306 ride offers for NYC and TKY respectively.

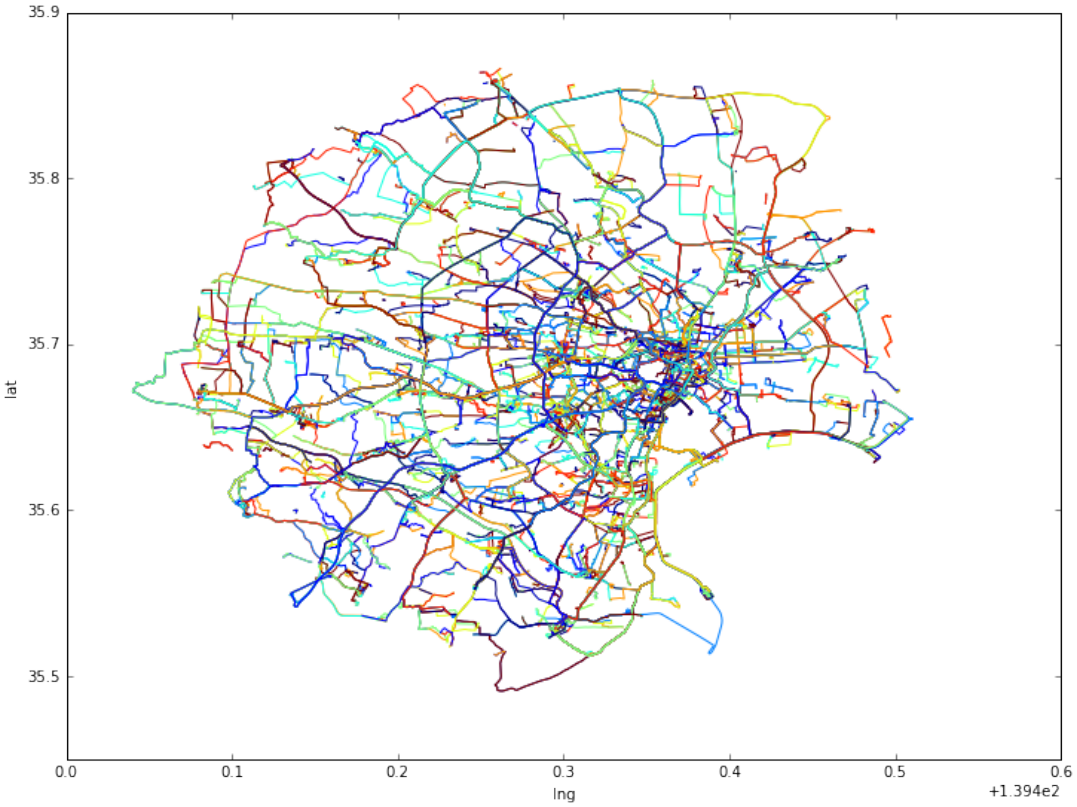
Figures 9a and 9b show a geospatial visualization of a sample of ride offers from the two datasets. Each color represents a single trajectory. As we can see, the ride offers cover the main streets and avenues of the cities, particularly in the downtown area characterized by a higher density.

For the extraction of the set Q of ride requests, we focus instead on the venues least frequently visited by each user. The insight is that occasionally visited venues are the ones for which a user is most likely to be open to accept a ride for an alternative destination. For example, venues like bars, restaurants, pubs, markets, cinemas are in general places not routinely visited for which we expect that users can be more flexible (LIRA et al., 2014). Based on this idea, we extract the ride requests for user u as follows. We first remove from the check-ins of u the check-ins in the two most frequently visited venues (see the previous procedure) and all the check-ins in venues belonging to the following categories: 'Home (private)', 'Office', 'Airport', 'Subway', 'Neighborhood', 'Road', 'Building', 'Residential Building (Apartment / Condo)', 'Government Building', 'Train Station', 'Road', 'Bus Station', 'Hotel', 'City' and 'Bridge'. We assume in fact that the activities associated to these categories of POIs can hardly be performed in alternative places. For example, a passenger could not be dropped to a different airport from the one she has the flight, or to a different hotel. Then, for each remaining check-in c we create a ride request q considering as starting location the place most frequently visited by the user. The

[4] developers.google.com/maps/documentation/directions/



(a) NYC



(b) TKY

Figure 9 – Samples of ride offers extracted from the FourSquare datasets.

destination is obviously the same of the check-in and it is associated to a POI category, while the departure time is inferred from the check-in time and the travel time needed to reach the destination from the starting location (as estimated by Google map). Resulting ride request corresponding to rides shorter than 1 km are discarded.

In addition, unless differently specified, we set in all the ride requests 60 minutes and 500 meters as the maximum delay and the maximum walk distance, respectively. We also consider 5.0 km/h as an average walking speed (MOHLER et al., 2007). These parameters allow us to represent different scenarios for ride sharing considering the extreme cases where the passenger waits from 1 minute up to 60 minutes and needs to walk from 0 meters up to 500 meters in total until reach his final destination. The resulting datasets contain 98,008 and 160,271 ride requests[5] in NYC and TKY, respectively.

4.4.2 Evaluation Metrics and Baseline

The acceptance of a ride offer is an absolutely subjective decision whose modeling is out of the scope of the present work. Thus, given the set of ride requests Q and the set of ride offers RT in the NYC and TKY datasets, we assess our proposal by simply considering the potential matchings returned by ABRM and the destination-oriented baseline, if any. In order to minimize the effect of differences in the implementations and ensure a proper analyse of the gain obtained by the proposed algorithm, the destination-oriented solution used as baseline is a modification of our implementation where the requested destination in the ride request is considered fixed in the matchings. Thus, similar to ABRM, the baseline also uses a pick-up time-window (as explained in Section 2.3.3.2), which define a maximum delay to pick-up the passenger, and uses a slugging approach, which consider that the passenger can walk to meet the driver at a pick-up point (as explained in Section 2.3.3.1). All the other parameters (including the maximum walking distance and time delay) are set exactly to the same values in order to directly measure the boost in the number of *requests potentially supplied* by our activity-based approach versus the destination-based counterpart.

Let us indicate with Q_s the subset of all ride requests Q satisfied by at least one ride offer. $|Q_s|$ is the number of ride requests potentially satisfied and $|Q_s|/|Q|$ the ratio (measured in percentage) between the number of requests supplied and the total number of requests in Q.

In order to estimate the potential impact of carpooling solutions on the reduction of kilometers traveled by cars, we assume ride requests are satisfied by the most highly ranked ride offer retrieved by ABRM and by the baseline, if any. We clearly assume a user can avoid to take its own vehicle when a ride possibility is offered. Every satisfied request

[5] The ride requests and POIs distributions by place categories of both datasets are reported in the Appendix, on Tables 17 and 18.

thus corresponds to one vehicle less in circulation.The number of kilometers potentially saved is thus:

$$Km_{saved}(Q_s) = \sum_{q \in Q_s} D(q.pickupLoc, q.dropLoc)$$

Where $D(q.pickupLoc, q.dropLoc)$ is the length of the route connecting the pickup and the (alternative) drop off locations estimated by the Google maps service.

4.4.3 RQ1: Can ride sharing opportunities be boosted?

In this section we address the first research question by comparing the number of ride sharing requests potentially satisfied by ABRM and the baseline.

We varied the method used by ABRM to choose the set of alternative venues. Specifically we experimented the following variations: (a) *all alternative destinations*, the passenger can accept to go to any alternative destination where she could perform the desired activity (all the venues in the same category of the requested POI); (b/c) *liked/popular*, the passenger can accept to go to one of the k most *liked/visited* POIs in the same category; (d) *preferred*, the passenger can accept to go to one among her k most preferred destinations. Since we do not have user-level preference information in our datasets, we simulate this case by randomly selecting k POIs in the same category for each user; (e) *Foursquare similarity*, the passenger can accept to go to one of the venues most *similar* to the requested destination according to the FourSquare similarity function (SEOL, 2015). Foursquare combines three important properties to compute the venue similarity: the co-visitation between the venue's visitors, the category and taste similarities between two venues. We remark that for the liked, popular and preferred criteria we experimented values of k equal to 5, 10 or 20. For the *Foursquare similarity* case, given the implementation of the APIs, at most 5 alternative POIs are returned even if in many cases a lower number of similar venues is suggested.

Table 3 reports the number ($|Q_s|$) and percentage ($|Q_s|/|Q|$) of ride requests potentially supplied, the improvement of ABRM compared to the destination-oriented baseline (gain), and the impact for the environment in terms of potentially saved kilometers (Km_{saved}).

As expected, we observe that the *all alternative destinations* method, due to the higher number of possible destinations, reaches the best results with a gain of 44.46% and 54.69%, compared to the baseline, for NYC and TKY, respectively. In general, we measured a higher performance in the Tokyo dataset compared to New York, probably due to the higher number of ride offers and alternative destinations. Another general trend we can note is that the number of requests potentially satisfied increases when more flexibility is assumed, i.e., when the number k of possible alternative destinations increases. On the other side we observed similarly that ABRM remarkably increases also

the number of ride offers matched to some ride requests. For example, in the NYC case, considering the most flexible criteria (i.e. all alternative destinations), 63% of the ride offers are matched to at least one ride request against 30% measured with the baseline.

Table 3 – Ride requests matched with the baseline and ABRM on the NYC and TKY datasets.

NYC	$\|Q_s\|$	$\|Q_s\|/\|Q\|$	*gain*	Km_{saved}
baseline	9,288	9.48%	-	35,901
all alt. destinations	52,864	53.94%	44.46%	388,293
likes (k=5)	18,378	18.75%	9.27%	117,027
likes (k=10)	24,534	25.03%	15.56%	157,840
likes (k=20)	32,483	33.14%	23.67%	213,254
popular (k=5)	20,268	20.68%	11.20%	129,702
popular (k=10)	25,011	25.52%	16.04%	160,634
popular (k=20)	32,614	33.28%	23.80%	213,908
preferred (k=5)	20,157	20.57%	11.09%	130,523
preferred (k=10)	28,627	29.21%	19.73%	190,249
preferred (k=20)	36,628	37.37%	27.90%	250,761
Foursquare similarity	12,642	12.90%	3.42%	55,783
TKY	$\|Q_s\|$	$\|Q_s\|/\|Q\|$	*gain*	Km_{saved}
baseline	36,695	22.90%	-	213,838
all alt. destinations	124,356	77.59%	54.69%	1,014,079
likes (k=5)	74,674	46.59%	23.70%	566,748
likes (k=10)	89,326	55.73%	32.84%	690,604
likes (k=20)	98,782	61.63%	38.74%	773,103
popular (k=5)	74,329	46.38%	23.48%	565,492
popular (k=10)	87,908	54.85%	31.95%	678,237
popular (k=20)	98,839	61.67%	38.77%	775,156
preferred (k=5)	67,044	41.83%	18.94%	513,269
preferred (k=10)	85,236	53.18%	30.29%	662,167
preferred (k=20)	97,223	60.66%	37.77%	764,810
Foursquare similarity	48,844	30.48%	7.58%	317,694

Table 3 also shows the estimated amount of kilometers traveled potentially saved by the corresponding carpooling solution. Once more we recall that the figures reported here are upper-bound estimates computed on the basis of the assumptions made on the flexibility and willingness of users to accept shared rides possibly at alternative destinations. From these values we can easily approximate the liter of fuel (e.g. gasoline) saved and consequently the saved amount of CO_2. For the sake of simplicity, in this computation we assume each car consumes in average 1 liter of gasoline for 7.449 Km as reported in

the official statistics of the Bureau of Transportation[6]. By assuming the passengers are flexible enough to change their requested location, the estimate for the number of kilometers saved in our test cities amount to 388.293 in NYC and 1,014,079 in Tokyo. These values in turn correspond to 52,120 and 136,117 less liters of gasoline and to 99.4 tons and 259.60 tons less CO_2 emission for New York and Tokyo, respectively. The baseline based on fixed destinations may save in NYC 35,786 Km only (corresponding to 4,803 gasoline liters and 1.22 CO_2 tons). For TKY we estimate instead that the baseline can save 213,838 Km (28,703 gasoline liters and 7.34 CO_2 tons).

On both the datasets, the method based on all the alternative destinations significantly outperforms the other methods. On the TKY dataset the average number of matches for each ride request is 690 for ABRM and only 3.02 for the baseline. Slightly similar figures are measured on the NYC datasets. The ride request in TKY with the highest number of alternative destinations counts about 23k matches against the 45 achieved by the baseline. Such high numbers motivate the need of a ranking model later discussed in the sections 4.4.4 and 4.4.5.

In conclusion, in relation to *RQ1*, we observe that the results reported prove the potential boost of ride sharing services involving the offers of rides to alternative destinations. The next step is to investigate how many these possibilities fit the passenger requests and the public good. These aspects are discussed in the next sections.

4.4.4 RQ2: How well do the ranked ride offers to alternative destinations meet ride request requirements?

In this section, we address our second research question, related to the contribution of the ride features to the ranking method defined in Section 4.2.1. With these features we intend to model how much the rides to alternative destinations can meet the requirements specified in the user request. Figure 10 supports this study. Each plot shows eight curves reporting the cumulative distribution of $|Q_s|/|Q|$ for the *baseline*, the *all alternative destination* criterion and the *popular* (k=5,10,20) and *preferred* (k=5,10,20) ones. The four plots for NYC (TKY) report each the effect on $|Q_s|/|Q|$ of varying the value of one of the features f_{dly}, f_{wld}, f_{dur}, and f_{len} by keeping all the other fixed.

We computed the cumulative distributions by considering the fraction of supplied ride requests having a value for the feature considered lower than the one reported in the x axis. In this way we can see how the fraction of supplied requests changes when the feature value increases.

Looking at the plots, specifically for the TKY dataset and the temporal shift feature f_{dly}, we see that even when we consider matches with only 10 minutes shift, more than half of the requests could be potentially satisfied with a ride offer to an alternative POI

[6] goo.gl/ClDSfL

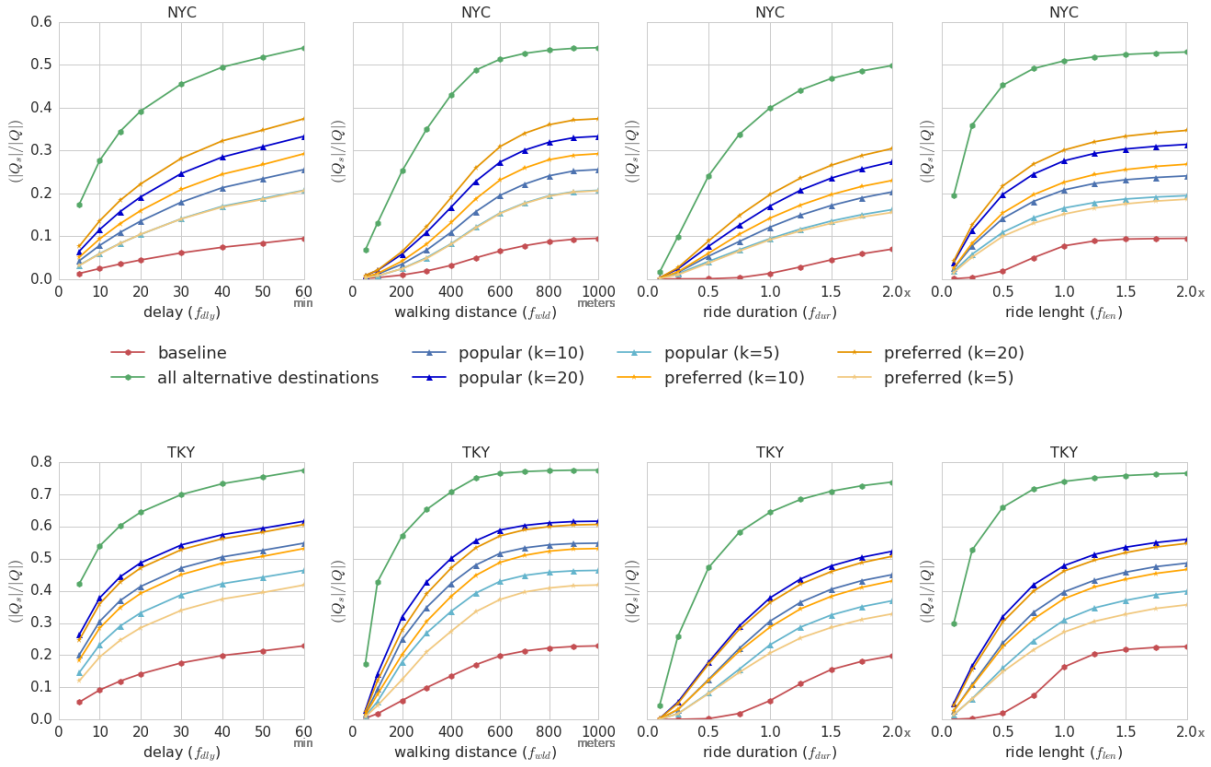


Figure 10 – Cumulative distribution of $|Q_s|/|Q|$ by varying each feature in isolation.

with the *all alternative destinations* method. A similar consideration can be done for the plot analyzing the variation on the walk distance feature f_{wld}.

For the trip length feature f_{len}, we report on the x axis the ratio between the matched ride offer having the shortest length and the shortest car distance to the requested destination computed by the Google Maps service. By looking to the plot corresponding to the NYC dataset we observe that more than 20% of the ride requests can be supplied to an alternative destination with a trip which is half in duration or shorter than the trip originally requested. We observe even better results on the TKY dataset. Similarly, for the the duration feature f_{dur} we report on the x axis the ratio between the estimated duration to reach the alternative POI and the travel time to arrive at the requested destination computed by the Google Maps service. On both the NYC and TKY datasets we can see that ABRM proposes rides to alternative destinations that are in most of the cases shorter in duration than the ones to the requested destination offered by the destination-oriented baseline. We can conclude this section by considering that independently from the feature considered, and thus from the subjective importance given by the user to the carpooling aspect modeled by the specific feature (time shift, walking distance, ride duration and ride length) ABRM is likely to provide a better carpooling service to flexible users accepting to reach alternative destinations for performing the intended activities.

4.4.5 RQ3: Effectiveness of the Ranking Method

ABRM returns the set of candidate ride offers matching a given ride request. The ranking model defined in Section 4.2.1 orders the candidate offers on the basis of a linear combination of weights. The purpose of the ranking step is helping the users to choose the most relevant rides among a possibly large set of offers. Since no golden standard recording user preferences is available to optimize the weights, we consider here a uniform weighting schema giving the same importance to all the features, and we compare such setting with four extreme scenarios where we prefer one feature over all the others by setting, in turn, the corresponding weight to 1 and the others to zero. Given the order of features previously used, we identify the uniform weighting schema with the vector $w = [1/4, 1/4, 1/4, 1/4]$ while the scenario giving, for example, only importance to ride duration (feature f_{dur}) corresponds to vector $w = [0, 0, 1, 0]$.

We chose the *all alternative destinations* as alternative destination criteria since it is the one providing the highest numbers of matchings, not biased by other parameters (like number of likes, popularity, etc.). As in the previous section of this chapter, f_{len} and f_{dur} are normalized based on the ratio between the estimated length and temporal duration to reach the alternative POI and the requested destination. For both datasets, the top-1 offer returned by the ranking model was compared with the best values of the extreme approach giving importance to only that feature. Figure 11 shows the cumulative distribution of value $|Q_s|/|Q|$ considering the top-ranked result for each of these five configurations of vector w.

As expected, the highest values for $|Q_s|/|Q|$ are achieved in all the plots for the extreme weighting schema considering the associated feature only. However, the curves corresponding to the uniform weighting schema are in all the plots, but the two in the most right hand side the closest to the highest curves, thus showing in general a very good performance. Uniform weighting performs as third solution under the considered metric only when the ride length aspect is analyzed, and the curves result to be very close to the second one. We observe in fact that the third and fourth features, ride duration and length, are highly correlated. On the other hand, the other two features, time delay and walking distance, are more selective on the generation of candidates. This suggests us that delay and walking distance are strong constraints to the matching of candidate offers, whereas the other two features may be weighted differently to try to optimize user acceptance on one hand and to improve public goodness by minimizing the total distance traveled by cars.

4.4.6 RQ4: Activities mostly favored by ABRM

We conclude the experiments by addressing RQ4, namely discussing which activities most favor the ride sharing in the alternative destination scenario. In other words, we analyze

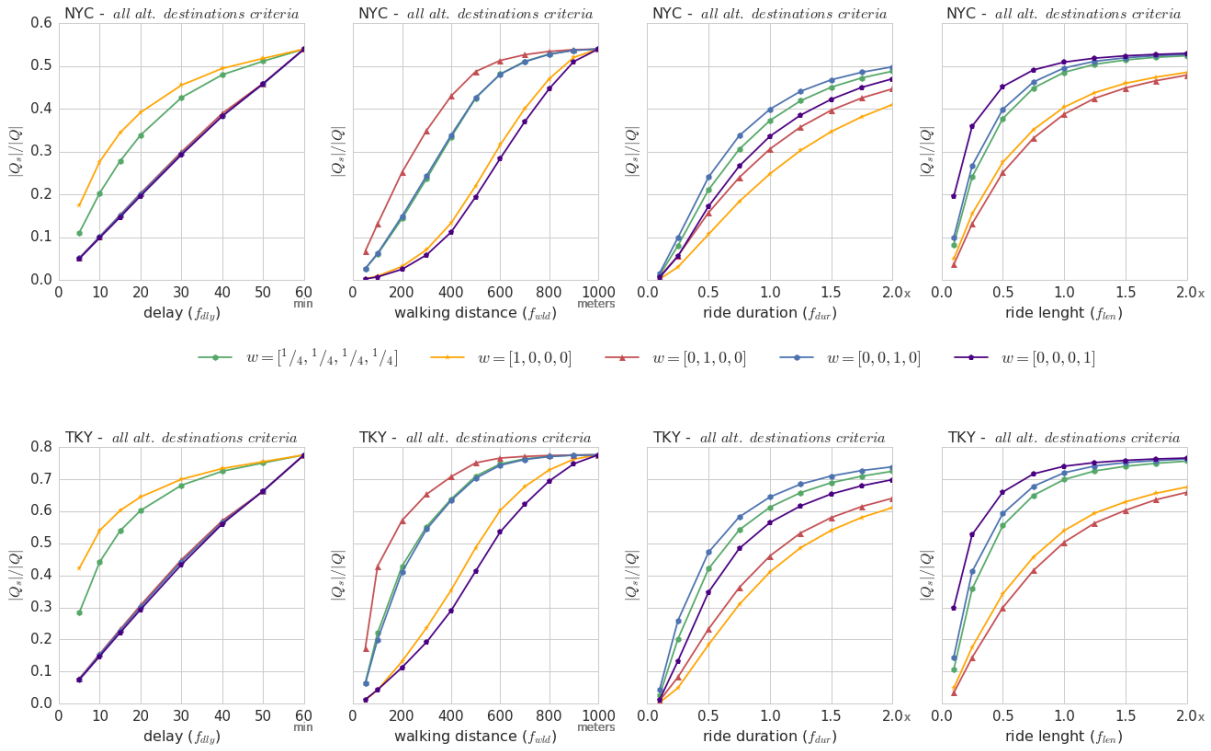


Figure 11 – Comparison by ride feature between the Top-1 ranked ride and the best result by feature

our ride matching results in the two datasets for understanding which activities have the larges boost in terms of number of ride matches.

We identified the most frequent ride requests by intended activity and, from them, we analyze the possible matches. Table 4 reports, for both datasets the top-10 activities that result in the largest number of ABRM ride matchings. We report the number of requests $|Q|$, the percentage ($|Q_s|/|Q|$) of ride requests potentially supplied with the baseline and the improvement in percentage obtained with ABRM. We use the *popular* with $k = 5$ (more restrictive) and *all alternative destinations* (less restrictive) as alternative destination criteria. We observe for NYC a boost on ride sharing possibilities mainly for activities related to "Italian Restaurant" and "Bar" with an improvement of +58.02% and +57.86% respectively. In turn, for TKY, activity "Food & Drink Shop" achieves the highest boost with +64.28%. This insight confirms our intuition that entertainment or eating are in general the activities that can most benefit from the proposed approach. The specific results reported in the table are however also correlated to a combination of factors such as the number of venues for each category in the cities and their location.

We thus conclude the experimental evaluation of the ABRM. The results have shown that the proposed algorithm increases the efficacy of the ride sharing systems for finding new ride matchings. Indeed, compared to the baseline, the ABRM improves up to 55% and 45% the number of ride request matched on the TKY and NYC datasets respectively.

Table 4 – Activities favoring ABRM boosting.

NYC	$\lvert Q_s \rvert$	Bas. $\lvert Q_s \rvert / \lvert Q \rvert$	ABRM *pop.*, $k = 5$	ABRM *all alt. dest.*
Italian Restaurant	1,315	13.08%	+12.97%	+58.02%
Bar	11,242	10.60%	+13.64%	+57.86%
Music Venue	1,027	9.45%	+25.61%	+55.50%
Coffee Place	918	8.06%	+20.59%	+54.68%
Chinese Restaurant	932	10.41%	+10.73%	+53.33%
Deli / Bodega	1,621	5.31%	+15.48%	+52.81%
Bakery	899	10.23%	+20.91%	+51.72%
Park	3,015	11.21%	+16.80%	+50.95%
American Restaurant	2,681	14.58%	+21.04%	+50.09%
Mexican Restaurant	1,468	13.42%	+19.69%	+49.11%
TKY	$\lvert Q_s \rvert$	Bas. $\lvert Q_s \rvert / \lvert Q \rvert$	ABRM *pop.*, $k = 5$	ABRM *all alt. dest.*
Food & Drink Shop	6,766	15.05%	+29.83%	+64.28%
Convenience Store	7,360	16.28%	+23.63%	+63.93%
Park	4,026	14.83%	+27.81%	+63.88%
Fast Food Restaurant	3,698	18.33%	+27.96%	+60.28%
Chinese Restaurant	2,804	20.97%	+29.81%	+60.06%
Japanese Restaurant	9,365	25.46%	+15.07%	+59.62%
Ramen / Noodle House	10,618	22.29%	+32.59%	+59.12%
Mall	6,185	20.03%	+26.89%	+57.83%
Coffee Shop	4,756	23.91%	+27.02%	+56.79%
Bar	8,051	26.89%	+33.21%	+55.21%

The following section introduces a demo developed upon the ABRM.

4.5 Comewithme - Demo Application

In this section, we present a demo application called ComeWithMe. ComeWithMe has been designed upon the ABRM Algorithm. For this reason, this ride sharing system is able to enlarge the candidate destinations of a ride request by considering alternative places where the desired activity can be performed. Activity-oriented carpooling hugely increases the number of rides matching a query, thus introducing requirements on system responsiveness and ranking effectiveness that are not common to traditional carpooling services.

4.5.1 Ride Search Engine

The main task of the *ride search engine* is to answer ComeWithMe passengers' queries by providing lists of ride candidate ranked according to the user context and preferences. This task is accomplished by mean of two important submodules that implement the *Retrieval* phase aspects of the ABRM Algorithm: *Query Expansion* and *Ranking Model*.

Query Expansion. This module boosts the possibilities of car rides by exploiting a query expansion technique. The use of query expansion generally increases recall and it is widely adopted in many application fields (CARPINETO; ROMANO, 2012). For this demo, the queries are ride requests expressing the passenger's intention to move to a venue to perform an activity. Given the destination *Point of Interest* (PoI) specified by the passenger, the query is automatically expanded with places related to the same activity by using a hierarchical thesaurus (an example is shown in Figure 12). The specific PoIs are the narrowest terms, while the intermediate layers represent different activities abstraction levels and thus possible query generalizations. For example, looking at Figure 12, when a passenger requests as destination "Da Gino", we see that it is an Italian Restaurant and expanding the query over Italian restaurants we have "Ristorante Giannino" as an alternative destination. Abstracting again up to "Eating" we have all the venues where they serve food corresponding to “Pizzeria”, “Japanese restaurants”, etc. The more we expand the query to broader terms, the more rides possibilities the passenger can select from the driver offers.

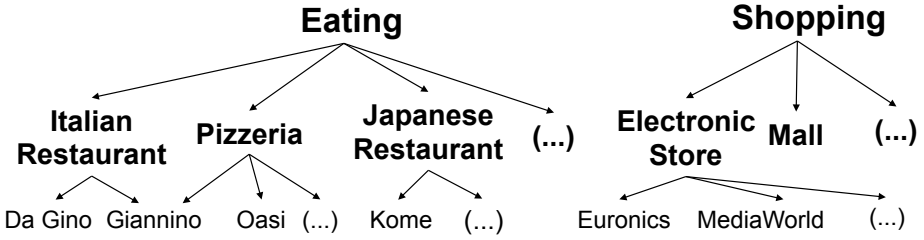


Figure 12 – Structure of the thesaurus.

Each venue in the thesaurus is associated with a cell of the spatial grid indicating

its location. Analogously, user queries are coded with the cells representing the pick-up area, the destination place, and a set of other cells representing alternative destinations. An example of the expansion process is illustrated in Figure 13: the destination PoI "Da Gino" is expanded with other possible venues (and cells) where the passenger can perform the activity "Eating".

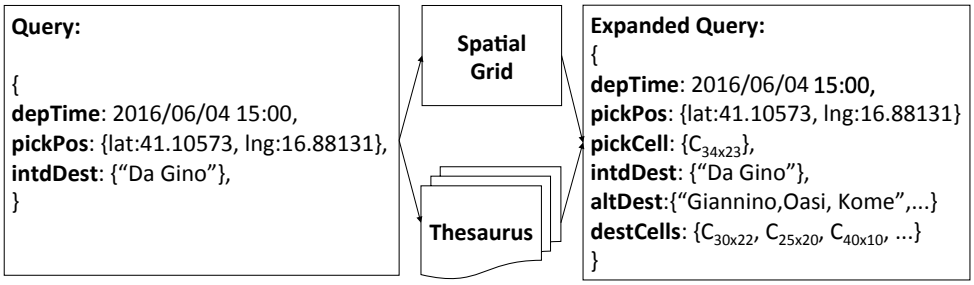


Figure 13 – A query expansion example having Italian restaurant "Da Gino " as destination place is expanded into a list of cells containing alternative places for eating.

Ranking Model. The ranking score of candidate rides is thus computed as a linear combination of a set of features, mainly derived from the flexibility preferences the passenger can set through the mobile app: 1) a temporal tolerance indicating the delay of the departure time of the trip respect to the preferred time indicated in the query; 2) a temporal tolerance on the possible anticipation of the trip respect to the indicated preferred time; 3) a spatial tolerance indicating how much the passenger is willing to walk to reach the pick-up point and/or the destination location. Other information considered in the computation of the ride score include the trip duration and the semantic similarity between the actual destination of the ride and the one specified in the query. Intuitively, the duration of the trip should not be too long respect to the duration of the fastest of all the possible rides. On the other hand, the destination venue should be, in order of preference: close to the PoI chosen in the query; another PoI in the same thesaurus category (e.g., a different Italian restaurant when the requested venue was an Italian restaurant); a PoI in the more abstract category of the thesaurus.

4.5.2 The Mobile Application

ComeWithMe has two different profiles of users: the driver, which offers rides, and the passenger which seeks for rides. The ride offers were extracted by a dataset of real car trajectories collected in the Tuscany regions. The database corresponds to a set of data with 44.278 trips, made by 5.048 users moving by car in the Tuscany. This dataset has been provided by OctoTelematics[7] company which installed a GPS device on cars for an insurance company.

[7] https://www.octotelematics.com/

The passenger interface allows the user to do ride request for a given destination, as shown in Figure 14. Once the required information is filled in, the user can submit the query and see the ranked list of rides offer.

In Figure 14, we see on the left our query example representing a user asking for a ride in the city of Pisa to go to the "Bella Napoli" pizzeria located in "via del Borghetto". The query specifies also the temporal tolerance (delay 30 minutes or anticipate 30 minutes) from the desired departure time at 19.00 and the spatial tolerance indicating the maximum distance the passenger is willing to walk (up to 600 meters).

ComeWithMe returns, for each query, a ranked list of rides where the best options are shown on the top. In our test dataset, during the specified temporal window (from 18:30 to 19:30), we have a total of 276 routine trips, 23 of which spatially matches the query from the pickup point to at least one "Pizzeria" among the 121 in the dataset. Since each trip can pass through many cells where "pizzeria" places are located, the ride search engine retrieves and ranks a total of 156 rides to "pizzeria" alternative destinations (see Figure 14 on the right). Observing the results of the query, we notice that the first two rides are to the intended pizzeria "Bella Napoli", while other destination are "Panuozzo" and "La Greppia".

The passenger can select a ride from the ranked list and visualize some information about the driver and other details about the ride (e.g. the pickup address and time, the destination place, the estimated arrival, etc). Once the user selected and confirmed a ride, ComeWithMe notifies the driver about the upcoming request. Symmetrically, as shown in Figure 15, a driver can see the list of passenger requests and she can select one to visualize the details. From the details interface the driver can accept or decline the request, she can call the passenger, start a chat and visualize the trajectory on the map.

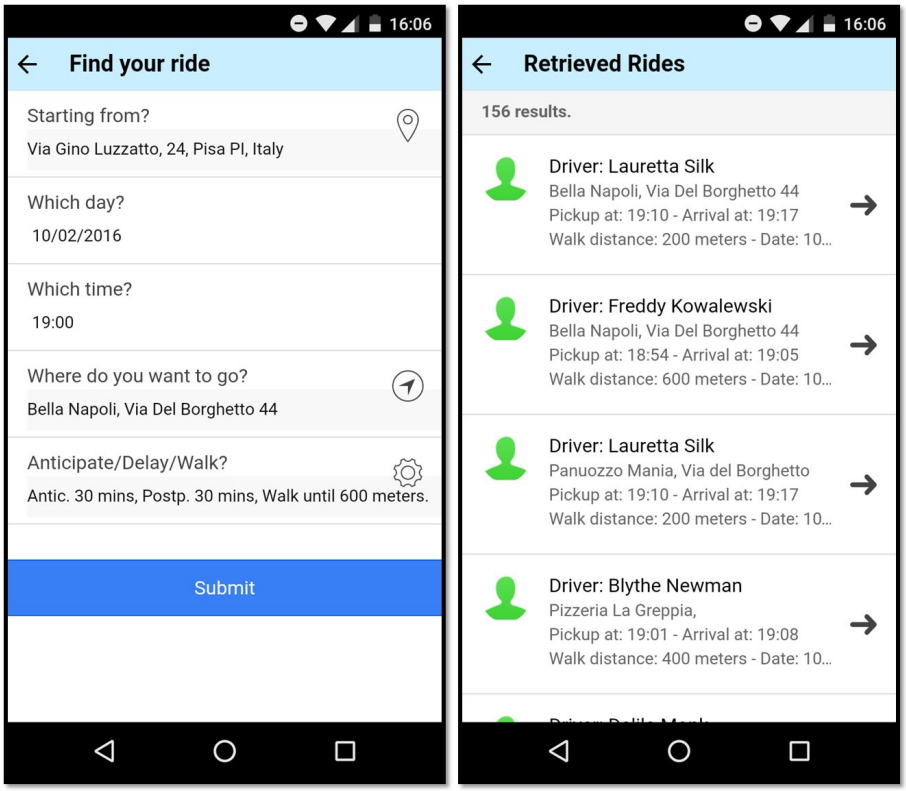


Figure 14 – Passenger Interface.

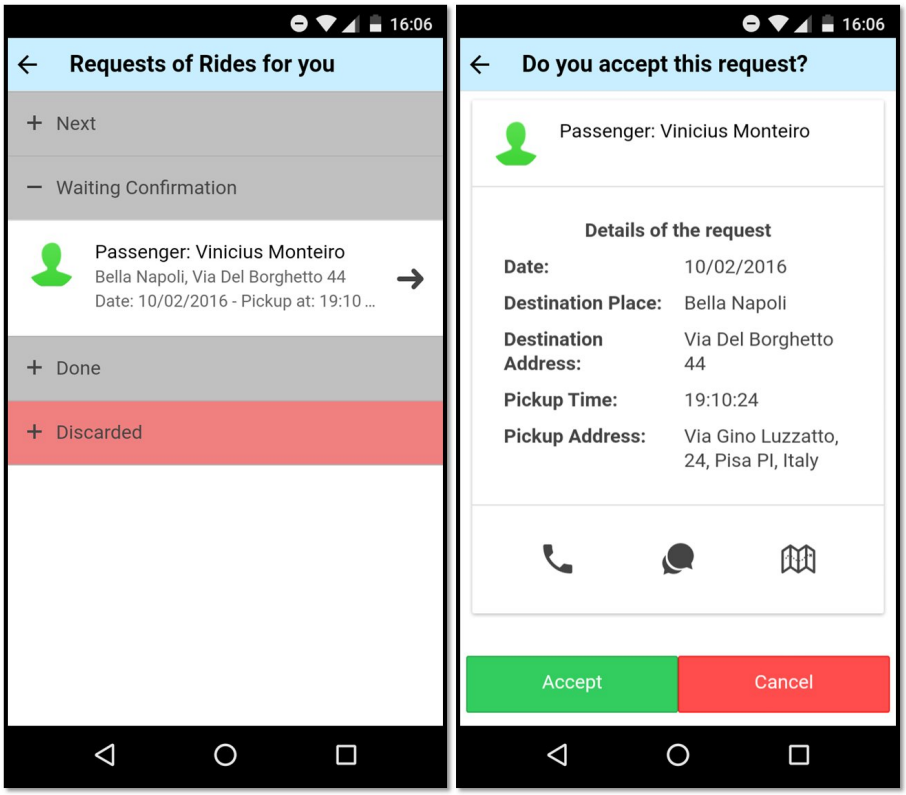


Figure 15 – Driver Interface.

4.6 Final Considerations

This chapter introduced and discussed the *Retrieval* phase of the Activity-Based Ride Matching algorithm (ABRM). The ABRM algorithm is aimed at matching ride requests with ride offers reaching alternative destinations where the intended user activity can be performed. Experiments conducted on two large semi-synthetic datasets recording mobility demands and the the categories of POIs visited (extracted from Foursquare checkins and made publicly available to favor the reproducibility of our results) showed that ABRM can boost to up 54.69% the percentage of ride request satisfied with compatible ride offers with respect to traditional destination-oriented ride sharing. Since the number of ride sharing opportunities provided by ABRM can be very large we proposed and analyzed in detail how the candidate ride offers can by ranked in order to better meet user expectations or to enforce their pro-environment behaviors in order to maximize the beneficial impact of carpooling on the environment in terms of CO_2 emissions saved.

Furthermore, we introduce ComeWithMe, which is a carpooling system offering alternative destinations designed upon the ABRM algorithm, thus boosting the number of available rides. The proposed prototype has been implemented to meet these features showing promising results.

5 Inferring transportation demands for large events using social media

Large event such as important music festival, or sports matches, motivate thousands of people to go to a specific location at a given time interval. A large event in fact requires a careful transportation planning to facilitate the attendees' arrival to the event's location. For such reason, often the event organizers provide to the participants dedicated transportation services as van or bus shuttles to supply the demand of rides.

By using social media, a common way of inferring the presence of users at events is considering the location associated with their media posts: the geotag, or "check-in", indicates the user presence at the time of the event at the event location. We observe, however, that this approach suffers from two drawbacks. The first drawback is that a small number of social media users enable the geotagging of their posts (in Twitter the percentage of geotagged posts is reported at about 2% (LEETARU et al., 2013; SLOAN et al., 2013)). Geotagged media posts represent a very sparse data source. For this reason, learning attendance prediction classifiers based on sparse data becomes extremely difficult. Another limitation is that the geotagged posts give as a clue about the positive attendance cases only, other means to derives the negative cases would be necessary. The second drawback of using only geolocated posts is that they represent the actual presence of the user at the event but not the *intention* of the user to participate in the event. Thus, the classifiers built using such sparse data would have a limited capacity to generalize and to infer the attendance of the users before or after the event. Indeed, we aim at predicting not only current attendance, but also the user participation in the event before and after the event.

Thus, given the attention to popular events reflected in social media, we want to classify user posts discussing an event on the basis of the actual attendance of the user to the event to enable or enhance several practical applications, not only in ride sharing, but also, for example, of targeted advertising and mobility management. From this analysis, we want to derive the key point of our investigation: *"Is it possible to infer past, current and future user attendance to large events through posts on social media to forecast the demand of rides?"*. Therefore, by inferring the future attendance, we can predict the users who will attend the event and potentially will need transportation services to reach the event's location. While, by inferring current and past attendance, we want to be able to understand who were the users who moved to the location of the event. These two latter subsets can support future transportation planning for the next editions of the event. The analysis of past transportation demands is a valuable input for transportation planners.

In order to present the proposed investigation this chapter is organized as follows:

Section 5.1 introduces our approach for classifying attendance and the features used to train suitable classifiers. Section 5.1.3 discusses the research questions addressed in this topic. In Section 5.2 the accuracy of each classifier is reported and analysed. In Section 5.3, we provide an example application of the deployed classifier for transport planning, while Section 5.4 provides concluding remarks.

5.1 Classifying Event Attendance

In the real-world, an *event* is something that occurs in a certain place during a particular interval of time. The location where the event occurs can be associated with its geographical coordinates (<lat, long>), while the temporal duration, which may vary from minutes to days or weeks, can be represented by a time window between a start time t_{start} and an end time t_{end}. In this work, we are interested in large events with thousands of participants. It is customary that such events have an associated entity in the most popular social media platforms (e.g. a Twitter account, a Facebook page), as well as a way of identifying discussions about them through the mentions of one or more *event identifiers* $i_1, \ldots i_n$, e.g., the event name, its acronym, some official or popular hashtags, etc.

A social media *post* by a user u, may contain text, links, emoticons, photos and/or videos (depending on the specific social network), as well as the timestamp at which the post was created and a social component representing the relations of u with other users (likes, followers, retweets, etc). In addition, some social networks permit the optional enrichment of the post with geotags, giving the <lat, long> position of the user when the post is made.

We define an *event-related post* p as any post that mentions one or more event identifiers and is thus possibly related to the specific event being considered. We distinguish these event-related posts as occurring *before the event* – when posted in a date before t_{start}, *during the event* – when posted between t_{start} and t_{end}, and *after the event* – when posted after t_{end}. Hereinafter, we will simply use the generic term *posts* to refer to event-related posts.

Our intuition is that the nature of event-related posts from attendees differ depending on *when* the posts are created. For instance, posts created before the event may express the users' intention to participate, or their regret for not being able to attend the event or regarding ticket sales. In contrast, posts published during the event may contain brief live reports from the event itself by the participating users, while non-attendees may express regrets for not being there, or comments about the coverage of the event on traditional or social media channels. After the event, attendees may share their opinions about the event, for example wishing to return to the event soon, while non-attendees may hope to participate in the next edition of the event. In Section 5.1.1, we illustrate these behaviors by providing some real-world examples of event-related posts. Later, in Section 5.2.4, we

validate these behaviors by analyzing the expressions most commonly used by users to positively or negatively convey event attendance.

Our work aims at understanding if these weak and noisy expressions of interest occurring in event-related posts can be exploited to identify the users who are likely to attend an event and distinguish them from those users that participate actively in the discussion about the event in social media but are not planning to attend it. In this last category we include user accounts directly linked to the event organization, as well as sponsors, advertisers and spammers. We propose to use supervised machine learning approaches to train binary classifiers that can automatically distinguish between posts of attendees and non-attendees. In order to consider the temporal dimension, we instantiate our attendance classification problem in three different tasks for the prediction of user attendance on the basis of posts published *before, during,* or *after* the date of the event.

5.1.1 Illustrating classification tasks Before/During/After the event

We argue that the types of posts made by users before, during or after an event tend to differ, and different classification models are necessary to attain an accurate classification of these posts.

Before Task: classifying attendance before the event. This task aims at predicting the attendance of a user at the event based on his or her shared posts at a time before the event. The classifier in this case exploits the content of posts where the users implicitly or explicitly express their intention to attend or not the event. Sometimes they explicitly share their intention to go with the words "Go" or "Packing" showing their intention to attend the event. Other common posts that might be considered as members of the negative class are those created by organizers, sponsors, or ticket sellers to provide general information about the event or advertisement and marketing material.

During Task: classifying attendance during the event. The aim of this task is to identify the users who, in the time window of the event, express their presence at the event. Very often, social media users express their actual participation in the event by posting photos or making comments about their experience during the event. On the other hand, non-attendees post general comments about their regrets for not attending or missing the event, or general comments without an explicit attendance meaning.

After Task: classifying attendance after the event. After the event is concluded, people often comment, express their opinions or publish memories and photos on social media. By inspecting such posts, it is often possible to obtain a clear determination of the user's attendance of the past event (positive) or not (negative).

Figure 16 shows some illustrative examples taken from our dataset related to a large UK music event (the Creamfields festival, see Section 5.2.1). From the content of

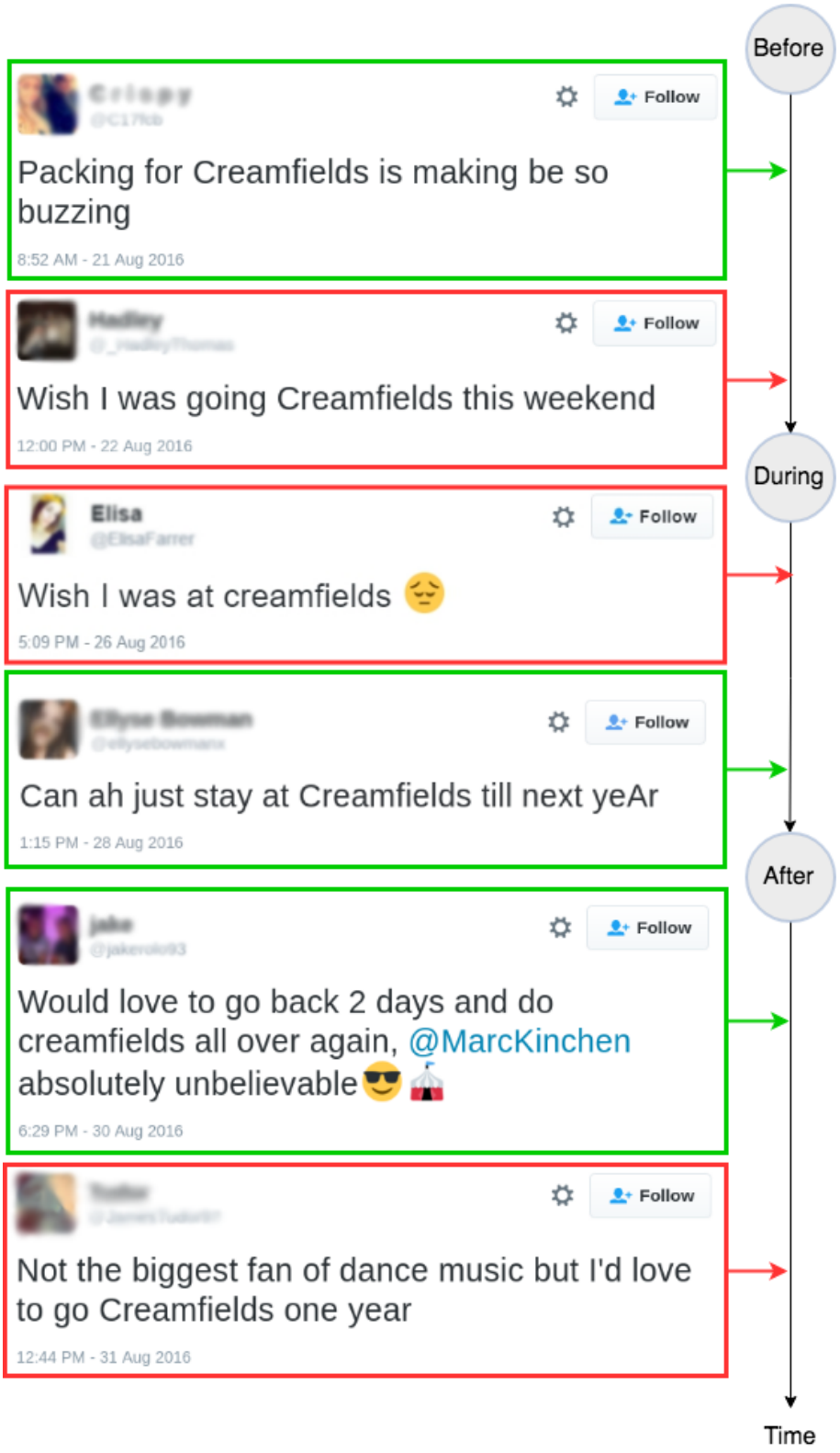


Figure 16 – Examples of tweets posted before, during and after the event.

the tweets reported in the figure, we can easily distinguish the positive (in green) and negative (in red) attendance cases for the *before*, *during* and *after* tasks.

5.1.2 Feature space for event attendance classification

We exploit four different categories of features. Each category reflects a different dimension of social media, namely the: textual, temporal, social, and multimedia dimensions.

- *Textual* features model the textual content of the post. We used two different methods for representing text. The first method uses a Bag of Words (BoW) model. In this case, the textual content is represented as the bag of unigrams, bigrams and trigrams occurring in the post. In order to reduce sparsity, we apply lemmatization to group together the different inflected forms of a word. Thus each lemma and each sequence of two and three adjacent lemmas are considered as features. Even if lemmatization reduces sparsity, still the BoW model cannot capture semantic relations among different lemmas. Let us consider for example a post with the words 'prepared to go' and another using the words 'ready to leave' instead. The same intention to attend the event is expressed in both the posts, but the BoW model does not capture such similarity. Later, we thus propose to encode the text in the posts by exploiting word embedding techniques based on word2vec (MIKOLOV et al., 2013). These techniques permit to reduce the dimensionality of the textual feature space and, at the same time, to capture text semantics. In addition to the previous features, we consider some additional features modeling textual metadata. Specifically, these features indicate the number of *words, hashtags, mentions, URLs* and *emoticons* occurring in the post. We discuss the text encoding techniques used and study the improvements achieved upon the BoW representation in Section 5.2.4.

- *Temporal* features represent the time of the post with respect to the event. The temporal dimension is needed to distinguish the classification task (*before*, *during* and *after*), but also to quantify how temporally distant from the event the post has been published. We simply represent time as the number of days separating the posting date from the event date(s). Such temporal feature is obviously meaningful only for the *before* and *after* classification tasks.

- *Social* features characterize the social profile of the posting user. Our social features include the number of followers, the number of followees and the ratio between them. An insight here is that users with a high number of followers and a relatively low number of followees are typically sponsors, organizers or VIPs that may advertise the event but do not necessarily attend it. Normal users targeted by our attendance classification task are indeed generally characterized by a lower number of followers and a more balanced followers/followee ratio.

- *Multimedia content* features identify whether a post has any multimedia content, such as a photo, video or a link to any visual content posted in other social networks such as Facebook or Instagram. Indeed, this feature group is motivated by the fact

that attendees may express their actual or past participation by posting photos or videos during and after the event. In addition, we observe that sponsors commonly use multimedia content before the event as a marketing tool.

It is worth noting that, in order to generalize the classification models learned, we removed the event identifiers $\{i\}$ from the textual content of all of the posts. The generalization aspect of our classifiers is studied in Section 5.2.2.2.

Table 5 – Features used split by category.

Textual	**Temporal**	**Social**	**Multimedia**
unigram bi/tri-grams			*num*:photos
num:words	*num*:days before	*num*:followers	*num*:videos
num:hash	*num*:days after	*num*:followees	*bool*:Youtube
num:mentions		*ratio*:(*num*:followers,	*bool*:Facebook
num:URLs		*num*:followees)	*bool*:Instagram
num:*emoticons*			*bool*:Foursquare
word embeddings			

Table 5 summarizes the features used by our classifiers grouped using the above four categories. The word embedding features are detailed in Section 5.2.2.2.

5.1.3 Research Questions

The overall aim of this work is to classify social media posts, shared by users before, during and after an event, as indicative of attendance or not attendance. We detail this classification objective into three tasks depending on the temporal aspect of the post: before, during and after. We study the behavior of the approach and, specifically, of the three classifiers, driven by three research questions. These questions will be answered in a number of experiments presented in Section 5.2. The research questions that we tackle are the following:

RQ1: *How accurate are our event attendance prediction classifiers?* This research question is discussed in details in Section 5.2.2 where we describe the accuracy results obtained by training supervised machine learning algorithms on an annotated dataset of media event-related posts. We will compare the obtained results with one baseline and discuss the performance achieved on the three different classification tasks. We introduce and discuss three more methods to improve the obtained accuracy. First, in Section 5.2.2.1, we conduct a *feature ablation* study to identify the feature groups that most contribute to attain high prediction accuracy. We will discover that the textual features are the most important, especially for the *before* and *after* tasks. This drives us to the study of word embedding as a way to reduce and enrich the feature space for this group of features. Section 5.2.2.2 discusses the improvement attained thanks to the word2vec encoding of post

texts. Finally, we conclude the study of RQ1 by assessing in Section 5.2.2.3 the accuracy of the classifiers on a further, objective, ground truth built by considering geo-located tweets.

RQ2: *How do these obtained classifier models generalize across events?* The possibility of deploying an event attendance classifier even when training data for the specific event is not available is highly desirable. In fact, some events do not have a large representation in social media or the cost of building a new training dataset could be unaffordable. The ability of our classifiers to generalize across events is thus of great importance. This research question is discussed in Section 5.2.3 where we assess how our models generalize across events by applying the model learned on one event to the other and vice-versa.

RQ3: *What are the most meaningful expressions posted by users to express their attendance to a given event?* This question is examined in Section 5.2.4 where we discuss the results of our analysis of co-occurrence and frequency of the most common terms in the posts classified as "attendance" or "not attendance".

5.2 Experimental Results

We instantiate our attendance classifiers in a scenario that considers two very popular music festivals held in the UK. Before addressing RQs 1-3, we first describe the setup of our experiments.

5.2.1 Experimental Setup

Our experiments are conducted using Twitter posts about two premier UK music festivals: Creamfields 2016 (held in Daresbury, UK, on August 25th-28th), and VFestival 2016 (held in Chelmsford/South Staffordshire, UK, on August 20th-21st). These events are notable in their size, with Creamfields in particular attracting over 70,000 attendees in 2016, and hence likely to be well-reflected in social media. Usually people publish event-related posts using specific hashtags and/or terms that refer to the event. We thus collected tweets related to these events by using the Twitter APIs for selecting tweets including the terms 'vfest' or 'v21st' and 'Creamfields'[1]. Tweets generated by the official accounts of the events (@vfestival and @Creamfields) were removed from the collections, since they are not relevant for our tasks.

For each respective event, the collected tweets are split on the basis of their timestamp into three different disjoint sets: *posts made before, during or after the event.* To

[1] Specifically, in order to cover the time periods before, during and after the considered events, we used the Twitter Streaming APIs from August 10^{th} to September 15^{th} 2016. Moreover, we used the Twitter REST APIs to collect the available tweets related to the events posted from March 1^{st} to September 15^{th} 2016.

Table 6 – Creamsfield and VFestival datasets statistics.

Dataset	Task	Labeled tweets	pos%	neg%	Tweets	Users	Geo-located tweets
Creamfields	**Before**	460	48.3	51.7	24,963	11,700	164
	During	460	39.1	60.9	25,625	15,884	309
	After	460	69.3	30.7	29,801	17,850	425
VFestival	**Before**	460	47.6	52.4	10,754	6,513	2
	During	460	37.4	62.6	4,873	3,285	75
	After	460	67.2	32.8	26,027	14,744	58

generate our training set, we randomly sample (without replacement) 460 distinct tweets for each task from each dataset, thus 1,380 tweets in total for each festival. Then, for each of the three tasks, a binary label is assigned to each tweet (positive class: a user who intends/is/has attended, and vice-versa for the negative class). The labelling task has been performed by a single assessor to keep the process consistent. On the other hand, we are aware of the limitations and risks of such human labelling process. In our specific case, we fortunately had the possibility of objectively validating the accuracy of our classifiers and the correctness of the adopted labelling procedure on a second, objective ground truth built from posts of georeferenced users. This analysis is reported in Section 5.2.2.3.

The human assessment is based on the textual or visual content of the tweet, which allows to establish any explicit evidence of attendance at the event. Any other kind of interpretation (advertisement, announcements, newsletter, sponsor's posts, sale of tickets, general information, regrets or impossibility, etc.) is labeled as negative. Table 6 reports for each dataset and task the number of labeled tweets, the respective percentage of positive and negative labeled tweets, the total number of tweets collected, and the number of distinct active users.

Specifically, we collected the tweets by geo-located users posted during the time window of the event and within an area of 3 km radius from the center of the event, gathering a total of 309 tweets from the Creamfields dataset and 75 tweets from the VFestival dataset. These tweets correspond to positive cases of attendance for the *during* task. Starting from these geolocated tweets, we identified a total of 189 distinct users for Creamsfield and 57 unique users for Vfestival who posted those tweets. We also gathered the event-related tweets posted by these users before and after the events. For the Creamfields event, we have 164 tweets before the event and 425 tweets after the event. For the VFestival dataset, we have 2 tweets before the event and 58 tweets after the event. All these tweets are included in a second test set as positive cases of pre- and post-events attendance. Table 6 summarizes in the 'geo-located tweets' column the number of tweets collected for each task by following the above procedure.

Our experiments are conducted using a 5-fold cross validation, while preserving the proportion of positive and negative instances in each fold. For each task and dataset, we train five different classification models, namely: Logistic Regression (LR), Gradient

Boosting Decision Trees (GBDT), Random Forest (RF), Support Vector Machine (SVM) and Naive Bayes (NB). All these algorithms, chosen among those that consistently delivering state-of-the-art performances in text classification tasks (AGGARWAL; ZHAI, 2012), are available in the scikit-learn library[2] used to train the classifiers. We use a grid search to tune the hyperparameters of the algorithms (BERGSTRA et al., 2011). Specifically: For LR, we consider L1 and L2 regularization and sweep the penalty parameter C in the range of {0.01,0.1,1,10,100,1000}; For GBDT and RF, we vary the number of trees in the range of {50, 80, 100, 120, 150}, while the learning rate and maximum tree depth vary in the ranges of {0.01, 0.05, 0.1} and {2,3,4,5}, respectively; For SVM, we use the RBF(Radial Basis Function) kernel with γ varying in {0.0001, 0.001, 0.01} and C in {0.01,0.1,1,10,100,1000}. For the vectorization and lemmatization of the textual content of the posts, we have used the scikit-learn library and the Natural Language Toolkit[3].

In the following, we report the performances achieved by our classifiers. Given that the classes are well-balanced in our labeled tweets, and for the peculiarities of the problem addressed both false positives and false negatives have a similar importance, we focus our analysis on classification accuracy values, which directly measure the number of correct predictions made divided by the total number of predictions made. For every classifier, we thus use the setting of hyperparameters that maximizes accuracy by using cross validation. Initially, we report accuracy, precision, recall, F1 and AuC for all classification models trained with the BoW text features. Afterwards, since, as we will show, the LR and GBDT classification models consistently outperform RF, SVM and NB, for the other experiments conducted, we report only the classification accuracy attained using these two classification approaches. The results of this work have been published in (LIRA et al., 2017; LIRA et al., 2019). The complete results of all classification models used in the experiments are reported in the Appendix of this thesis.

5.2.2 Results: RQ1

In this section we address RQ1 - studying the accuracy of our event attendance prediction classifiers.

Table 7 reports the accuracy, precision, recall and F1 measure of our 5 classifiers on each dataset and classification task (*before*, *during*, *after*). For the classifiers reported in this table, all feature groups are used, with the textual content of posts represented according to the BoW model. On analysing the results in Table 7, we find that our GBDT classifiers attain the highest performance for all the tasks on the VFestival dataset with an accuracy and precision always greater than 80%. For posts made during the event, GBDT obtained an accuracy of ∼82% when classifying the attendance of the users at the Creamfields event and also when inferring past attendance at VFestival. The performance

[2] <http://scikit-learn.org/>
[3] <https://www.nltk.org/>

Table 7 – Classification effectiveness using BoW features.

	Dataset: Creamfields						Dataset: VFestival					
Task	**Model**	**Acc.**	**Prec.**	**Recall**	**F1**	**AuC**	**Model**	**Acc.**	**Prec.**	**Recall**	**F1**	**AuC**
Before	LR^{bow}	0.868	**0.870**	0.870	0.868	**0.887**	LR^{bow}	0.761	0.744	0.762	0.748	0.764
	$GBDT^{bow}$	**0.874**	0.846	0.912	**0.878**	0.873	$GBDT^{bow}$	**0.809**	0.802	0.768	**0.784**	**0.808**
	NB^{bow}	0.587	0.540	**0.977**	0.696	0.600	NB^{bow}	0.535	0.506	**0.977**	0.667	0.555
	RF^{bow}	0.826	0.760	0.941	0.840	0.830	RF^{bow}	0.778	**0.860**	0.648	0.735	0.772
	SVM^{bow}	0.607	0.591	0.599	0.593	0.606	SVM^{bow}	0.578	0.568	0.471	0.514	0.573
During	LR^{bow}	0.741	0.766	0.538	0.602	0.690	LR^{bow}	0.626	0.600	**0.614**	0.494	0.606
	$GBDT^{bow}$	**0.817**	**0.830**	**0.616**	**0.708**	**0.790**	$GBDT^{bow}$	**0.802**	**0.850**	0.582	**0.688**	**0.763**
	NB^{bow}	0.628	0.619	0.117	0.193	0.537	NB^{bow}	0.530	0.429	0.737	0.525	0.571
	RF^{bow}	0.620	0.600	0.028	0.053	0.514	RF^{bow}	0.680	1.000	0.145	0.248	0.573
	SVM^{bow}	0.641	0.584	0.300	0.394	0.580	SVM^{bow}	0.670	0.800	0.157	0.257	0.566
After	LR^{bow}	**0.813**	**0.810**	0.958	**0.880**	**0.762**	LR^{bow}	0.809	0.812	0.932	**0.868**	**0.808**
	$GBDT^{bow}$	0.780	0.792	0.948	0.864	0.640	$GBDT^{bow}$	**0.815**	**0.824**	0.902	0.862	0.767
	NB^{bow}	0.702	0.711	0.962	0.818	0.538	NB^{bow}	0.696	0.709	0.929	0.804	0.574
	RF^{bow}	0.713	0.708	**1.000**	0.829	0.532	RF^{bow}	0.689	0.684	**1.000**	0.812	0.527
	SVM^{bow}	0.707	0.706	0.991	0.824	0.527	SVM^{bow}	0.707	0.699	0.994	0.820	0.556

achieved with GBDT on the VFestival dataset for the *after* task is also good with an accuracy of nearly ∼82%. LR outperforms GBDT for all metrics on the *after* task at the Creamfields, while it attains a better recall in other two cases (*during* and *after* tasks for VFestival).

In summary, for RQ1, the accuracy results reported in Table 7 show that our approach is reasonably effective at classifying user attendance. We observe that GBDT on average outperforms the other algorithms and LR achieves the best accuracy in one of the six cases.

5.2.2.1 Feature groups that are most effective in attaining high prediction accuracy

In this section, we explore in more details the previous results by analysing the contribution of the feature groups defined in Section 5.1.2: multimedia, social, temporal and textual feature groups. Our objective is to understand which feature group deserves further study because it provides the largest benefit to attain a high prediction accuracy.

To evaluate the contribution of each group of features, we conduct an ablation study. Specifically, we remove each group of features one at a time from the datasets used to train and test the classifiers. For such analysis, we use the GBDT classifier, which, according to the results reported in Section 5.2.2, on average achieves the highest performance. Table 8 reports the results of the ablation study sorted by accuracy for each of the *before*, *during* and *after* classification tasks. In the table, each row denoted with 'All - *feature_group*' indicates that the features of group '*feature_group*' were ablated (removed).

On examination of Table 8, we find that the multimedia features are very important for the *during* task, particularly for VFestival, where a ∼5% drop in accuracy is observed when the multimedia feature group is ablated (0.802 → 0.757). Indeed, in this dataset, we

note that 0.85%, 22% and 27% of the tweets posted, respectively before, during and after the event have some multimedia content. For Creamfields, the corresponding percentages tweets containing multimedia content are 0.4%, 8% and 20%, respectively.

Next, we note that the social features are useful for the *before* task in Creamfields and for the *after* task on VFestival, where their exclusion implies a loss of accuracy. We note that these features allow to identify (negative) advertisement posts coming from event sponsors or news providers, all of whom have a high number of followers.

The temporal features are important when classifying attendance after the completion of the event. We note that low values for this feature (i.e. a shorter difference between the dates before or after the event) are indicative for identifying the actual attendees of the event, while higher values (i.e. more distant from the event) are indicative for identifying non-attendees. This is reasonable when observing the real-world, where people who participated in events discuss them on social media only for a short period of time, usually for a few days before or after the event. Sponsors and news providers, instead, tend to post about the event regularly over a longer time period for marketing purposes.

Indeed, by manually inspecting the posts in our training datasets for the tasks before and after, we found that more than 82% of the distribution of posts published by attendees is concentrated in a time interval of 5 days before and 3 days after the event, while the posts of sponsor accounts are more uniformly distributed over time.

Users express their attendance at an event through the post text in different ways depending on the period (before, during, after). Hence, as highlighted in Section 5.1.1, the textual features extracted from the posts vary depending on the task. As we can see from the table, textual features are the most important for the *before* and *after* tasks. For these tasks, in both datasets, once we exclude those features, the accuracy drops tightly. Furthermore, in all experiments, keeping the textual features allows the models to achieve good accuracies, close to the optimal cases. Before the event, the users mention often their participation by posting about the purchase and delivery of their tickets (feature 'ticket' is among the most important for both Creamfields and VFestival), or when they express their anxiety to attend the festival (e.g. features such as 'wait' and 'excited'). After the event, the users share their experience, how they feel after the event and state willingness to come back to the next edition.

Lastly, the meta textual content (number of *words, hashtags, mentions, URLs* and *emoticons*) only exhibit an importance for attaining accurate classifications for the *before* task of the VFestival. For the same festival and for the *after* task, these features introduce noise into the GBDT model, since the exclusion of this set of features marginally improves the accuracy of the model.

Finally, as a summary of our findings, we observe that while each of the feature groups has some impact for at least one of the tasks, we highlight again the usefulness of the textual features for the prediction of attendance for all the tasks. Indeed, when

Table 8 – Accuracies of the GBDT models by ablating groups of features.

	Creamfields		**VFestival**	
Task	**Group**	**Accuracy**	**Group**	**Accuracy**
Before	All	**0.874**	All	**0.809**
	All - Temporal	**0.874**	All - Social	**0.809**
	All - Multimedia	**0.874**	All - Textual_meta_feats	**0.809**
	All- Textual_meta_feats	0.865	All - Multimedia	0.794
	All - Social	0.863	All - Temporal	0.792
	All - Text	0.606	All - Text	0.656
During	All	**0.817**	All - Textual_meta_feats	**0.806**
	All - Textual_meta_feats	0.815	All	0.802
	All - Social	0.811	All - Text	0.802
	All - Multimedia	0.804	All - Social	0.791
	All - Text	0.667	All - Multimedia	0.757
After	All - Social	**0.793**	All	**0.815**
	All - Textual_meta_feats	0.787	All - Textual_meta_feats	0.811
	All	0.780	All - Temporal	0.809
	All - Temporal	0.780	All - Social	0.807
	All - Multimedia	0.769	All - Multimedia	0.781
	All - Text	0.689	All - Text	0.724

this group is ablated from the model, the classification accuracy decreases remarkably on both datasets. This observation suggests to attempt improving the results by enriching the group of textual features. This research direction is investigated in the next section.

5.2.2.2 Classification accuracy improvement from word-embedding features

In the context of RQ1, the analysis in this section aims to investigate new features that could enhance the performance of our classifiers. Thus far, in our models, the textual content of posts has been represented as BoW features. One drawback of BoW is that different words have different representations, regardless of their semantic meaning (BALIKAS; AMINI, 2016; MCDONALD; MACDONALD; OUNIS, 2017). For example, while the words 'buy' and 'purchase' have similar meanings (synonyms), in a BoW representation they are as similar as two antonyms. This is not desirable for our attendance classifiers that aim to capture the semantic of the users' posts. To tackle this problem, we use word2vec, a neural net learning technique that embeds words from a vocabulary into a vector space, which represents the linguistic contexts of words - namely, that words that have similar meanings are represented by close vectors in the embedding space. Specifically, we use the *gensim* [4] implementation of word2vec and a word2vec model trained on part of the

[4] <https://radimrehurek.com/gensim/models/word2vec.html>

Google News dataset (about 100 billion words)[5]. This model contains 300-dimensional vectors for 3 million words and phrases. We also conducted initial experiments using a word2vec model trained on a large Twitter corpus[6] (about 400 million twitter microposts). However, since the results of the experiments conducted using the Google News model slightly outperformed those with the model trained on the Twitter corpus, we report only the former in the following experiments.

We represent each post with a single 300-dimensional vector obtained by combining the vectors that represent all the terms occurring in the post. This combination can be done with different aggregation functions. We explore the use of the 'sum', 'mean' and 'max' aggregation functions and also the concatenation of these three representations that we denote by *mix*. The aggregation functions 'sum', 'mean' and 'max' have the intuitive meaning of building a single vector for a post by computing the sum (respectively, mean, max) among the 300 dimensions in the embedding of all the posted words. Differently, the *mix* representation of the post consists in simply using the concatenation of the above three aggregated vectors.

Table 9 reports the performances achieved by the GBDT and LR models trained: (a) using BoW features (denoted by *bow*); (b) using word2vec features (denoted by *w2v*) instead of BoW; (c) using both the BoW and w2v features (denoted by *both*). For these experiments the other groups of features (social, temporal and multimedia) are also included in the training sets. For the sake of simplicity, the table reports only the best results achieved by a given algorithm with the corresponding sets of features. For example, the notation $\text{GBDT}^{\text{both}}_{\text{mean}}$ means that the GBDT classifier is trained using *both* BoW and w2v features and that the w2v representation of the post is obtained using the *mean* aggregation function. Similarly, $\text{LR}^{\text{w2v}}_{\text{sum}}$ means that the LR model was trained using w2v features aggregated with *sum*.

We observe that, in general, the use of w2v features improves the classification accuracy compared to the sole use of BoW features (*bow*). Indeed, for the Creamfields dataset, the use of embedding features improves the accuracy and precision figures up to ∼91%. It is worth noting that the improvement in Accuracy is higher with LR. Indeed, when the w2v textual features are used, either jointly with BoW (*both*) or not (*w2v*), the LR classifiers improve by +4.5%, +7.9% and +2.6% the Accuracy on the *before, during, and after* tasks on the Creamfields, respectively. Further large improvements are achieved on the VFestival dataset where we observe +5.2%, +16.1%, +4.9% in accuracy for the three tasks. Moroever, the GBDT models attain increased accuracy when using the embedding features, although they are more remarkable for the *after* task. Here, we observe improvements up to ∼5% (0.78 → 0.833 on Creamfields and 0.815 → 0.861 on VFestival) when using only the w2v features. On closer inspection, we see that the w2v features

[5] <https://github.com/mmihaltz/word2vec-GoogleNews-vectors>
[6] https://github.com/loretoparisi/word2vec-twitter

enhance the classification accuracy almost independently of the tasks and algorithm used to train the model. Compared to the results using only the BoW features, the Accuracy is most increased for the *before* (0.874 → 0.913 for Creamfields) and *after* (0.815 → 0.861 for VFestival) tasks. In these tasks, as discussed above, the textual features have high importance for accurate classification, thus the embedding features provide meaning in a lower-dimensional space that allows for more accurate models compared to the other features.

Table 9 – Accuracy of the GBDT and LR classifiers trained with BoW, w2v and both(BoW+w2v) features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

	Creamfields			**VFestival**		
Task	Model	Accuracy		Model	Accuracy	
Before	LR^{bow}	0.868		LR^{bow}	0.761	
	LR^{w2v}_{mix}	0.885	(+1.7%)	LR^{w2v}_{mix}	0.778	(+1.7%)
	LR^{both}_{max}	**0.913***	(+4.5%)	LR^{both}_{sum}	0.813*	(+5.2%)
	$GBDT^{bow}$	0.874		$GBDT^{bow}$	0.809	
	$GBDT^{w2v}_{sum}$	0.874	(0.0%)	$GBDT^{w2v}_{max}$	0.818	(+0.9%)
	$GBDT^{both}_{mean}$	0.872	(0.0%)	$GBDT^{both}_{max}$	**0.824**	(+1.5%)
During	LR^{bow}	0.741		LR^{bow}	0.626	
	LR^{w2v}_{sum}	0.800*	(+5.9%)	LR^{w2v}_{mix}	0.772*	(+14.6%)
	LR^{both}_{mix}	**0.820***	(+7.5%)	LR^{both}_{mix}	0.787*	(+16.1%)
	$GBDT^{bow}$	0.817		$GBDT^{bow}$	0.802	
	$GBDT^{w2v}_{max}$	0.789	(0.0%)	$GBDT^{w2v}_{max}$	0.823*	(+2.1%)
	$GBDT^{both}_{mix}$	0.796	(0.0%)	$GBDT^{both}_{max}$	**0.826***	(+2.4%)
After	LR^{bow}	0.813		LR^{bow}	0.809	
	LR^{w2v}_{sum}	0.824*	(+1.1%)	LR^{w2v}_{mix}	0.850*	(+4.1%)
	LR^{both}_{sum}	**0.839***	(+2.6%)	LR^{both}_{sum}	0.858*	(+4.9%)
	$GBDT^{bow}$	0.780		$GBDT^{bow}$	0.815	
	$GBDT^{w2v}_{max}$	0.830*	(+5.0%)	$GBDT^{w2v}_{mix}$	**0.861***	(+4.6%)
	$GBDT^{both}_{max}$	0.833*	(+5.3%)	$GBDT^{both}_{mix}$	0.854*	(+3.9%)

5.2.2.3 Assessment of accuracy on the geo-located tweets

As a further evaluation of the classifier accuracy, we test the models with the second ground truth dataset composed by geo-located tweets. Recall that the fraction of geo-located tweets is very low, thus making any approach based on geo-location only not feasible for addressing our event attendance classification problem. However, since the geo-location confirms the presence of the user at a given place, we can exploit the geo-

Table 10 – Accuracy of the classifiers on the geo-located tweets.

	Creamfields		**VFestival**	
Task	Model	Accuracy	Model	Accuracy
Before	LR^{w2v}_{mean}	**0.854**	LR^{both}_{max}	0.500
	$GBDT^{bow}$	0.726	$GBDT^{w2v}_{sum}$	**1.000**
During	LR^{both}_{mean}	0.958	LR^{both}_{sum}	1.000
	$GBDT^{bow}$	**0.964**	$GBDT^{w2v}_{sum}$	**1.000**
After	LR^{w2v}_{sum}	0.934	LR^{both}_{mean}	0.844
	$GBDT^{bow}$	**0.960**	$GBDT^{both}_{sum}$	**0.879**

located tweets in our dataset to further assess the validity of our approach on a second independent test set having no intersection with the training set. In addition, this second experiment permits to indirectly validate the labeling procedure adopted to generate our ground truth.

Table 10 shows the performances of our best performing LR and GBDT models on this second test set. We measured very high accuracies, always higher than 85%, on each classification task. Accuracy reaches 96% and 100% on the *during* task for the Creamfields and VFestival events, respectively. Since the above classification accuracies are higher than those measured on the other test sets, we manually inspected the geo-located tweets in these second test sets. We observed that for both festivals, for the *during* task, about 90% of the geo-located tweets contain some multimedia content. The percentage of during posts including multimedia content in the original ground truth were instead much lower: 8% and 22% for the Creamfields and VFestival events, respectively. As discussed in Section 5.2.2.1, multimedia features are among the most important for the *during* task.

The high classification accuracy achieved on the georeferenced posts validates the correctness of the adopted labeling procedure. Finally, it strongly confirms the quality of our attendance prediction classifiers and the validity of our approach based on the content of tweets only.

5.2.3 Results: RQ2

Our second research question (RQ2) aims to determine how the classifiers can generalize to other similar events (in our case, music festivals). Indeed, while our experiments are conducted over two datasets representing two music festivals, these events have some specific differences. For instance, the VFestival event is a music festival for pop music, while Creamfields is an electronic music festival, with distinctly different genres of performing artists. Therefore, these events may attract different kinds of attendees and may lead to different discussions on social media, reflecting different ways of expressing attendance at the event.

In order to address RQ2, we conduct experiments by applying the model trained on one dataset to classify the labeled samples of the other dataset and vice-versa. The results of these experiments are shown in Table 11. We observe that our classifiers attain reasonable performances even across different events. The classifiers trained on the VFestival dataset achieve an accuracy ∼87% (LR^{w2v}) and ∼81% ($GBDT^{both}$) for the prediction of attendance before and after the Creamfields event respectively. Accuracy however drops to ∼75% ($GBDT^{both}$) for the *during* task. One possible reason for this drop is that for the VFestival training set the most relevant features for the classification during the event are the posts with photos and Instagram, while for the Creamfields dataset, the textual features were observed to be the most useful (as discussed in Section 5.2.2.1).

Table 11 also shows that the LR and GBDT models achieve the highest accuracies when using only w2v features or both BoW and w2v features. The table also shows the improvement of GBDT and LR compared to the use of only BoW features. As expected, we note that the word embedding features substantially boost the performance of cross-event classification with respect to models using BoW features only. Indeed, when training the models with the Creamfields dataset and testing it on VFestival for the *after* task, the GBDT accuracy goes from 71.7% with $GBDT^{bow}$ to 78.9% with $GBDT^{w2v}$ (+5.6% improvement compared to the $GBDT^{bow}$). Moreover, LR reaches 78.7% with LR^{both} w.r.t. 72.0% with LR^{bow} (+6.7 %). Answering RQ2, we can conclude that our classifiers, trained on one event and tested on the other, generalize well, particularly benefiting by the abstraction from the specific event provided by the use of w2v features.

Table 11 – Generalization ability of the classifiers: models trained on Creamsfields are tested on VFestival and vice-versa. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**Creamfields/VFestival**			**VFestival/Creamfields**		
Task	Model	Accuracy		Model	Accuracy	
Before	LR^{both}_{mix}	**0.796**	(+1.3%)	LR^{w2v}_{mix}	**0.865***	(+1.3%)
	$GBDT^{bow}$	0.780	(0.0%)	$GBDT^{bow}$	0.824	(0.0%)
During	LR^{both}_{max}	0.702*	(+3.2%)	LR^{both}_{sum}	0.741*	(+9.2%)
	$GBDT^{w2v}_{max}$	**0.724***	(+1.3%)	$GBDT^{both}_{max}$	**0.743***	(+3.2%)
After	LR^{both}_{mix}	0.787*	(+6.7%)	LR^{bow}	0.787	(0.0%)
	$GBDT^{w2v}_{sum}$	**0.789***	(+5.6%)	$GBDT^{both}_{mean}$	**0.807***	(+3.7%)

5.2.3.1 Improving the robustness of the classifiers.

We now conduct experiments to understand if the generalizability of our classifiers can be enhanced. In doing so, we use the annotated dataset to understand if a given term occurring in a post is more indicative of attendance or not attendance. To this end,

we count the occurrences of all the terms in the positive or negative posts of our gold standard, and consider the normalized frequency of the term in the respective classes as an indicator of whether a word is more likely to be associated with event attendance or not. For example, a term occurring 10 times in the gold standard, 4 times in posts expressing attendance and 6 times in negative ones, is scored 0.6 for attendance and 0.4 for not attendance. By aggregating (with 'sum', 'mean' and 'max') such values for each term occurring in a post, we can generate two additional features to be used for the classification tasks. Furthermore, the concatenation of the 'sum', 'mean' and 'max' representations is considered (denoted as 'mix'), generating then six additional features (i.e. two features for each aggregation). However, these values are available only for terms occurring in the training set and posts to be classified can include "out-of-vocabulary" (OOV) terms not in this set (KAEWPITAKKUN; SHIRAI; MOHD, 2014).

The word2vec features provide us with a solution to address the OOV issue. Specifically, given a term t occurring in a post but not present in the training set, we compute its embedding vector v and retrieve the top-k most similar vectors (using Cosine similarity (LEVY; GOLDBERG; DAGAN, 2015)) for which the feature is available from the training set. The feature for t is finally computed as the average of the features associated with the k closest vectors weighted by the cosine similarity. The intuition behind this idea is that terms with similar embedding vectors have also similar semantics. We indicate this approach as *Normalized Frequency Vectors* (*NFV*), and report the results of experiments where we varied the value of k in the range of 1, 3 and 5.

Table 12 reports the accuracy performances for the LR and GBDT classifiers exploiting the NFV features measured across the datasets. In the table, we report the improvement in accuracy achieved over the best results reported in Table 11 and the operators used for aggregating the embedding vectors and the NFV features.

From Table 12, we observe that the NFV features enhance the accuracy of our attendance classifiers up +2.4% and +3.5%, on the VFestival and Creamfields events, respectively. However, the *during* task still attains the lowest classification accuracies compared to the other two tasks. Furthermore, we see from the table that, for all of the tasks, the accuracy is higher when training uses the VFestival datasets, thus suggesting some overfitting of the models trained on the Creamfields data. In general however, for most of the tasks and models, we observe statistically significant performance improvements (McNemar's test, $p < 0.05$), corroborating our expectations of the usefulness of the NFV features for the robustness of the classifiers. To better understand how the context and semantic behind the embedding features can help the classification, we investigate in the next section how the semantic similarity among terms actually contribute to the robustness of the models.

Table 12 – Robustness of the classifiers exploiting the NFV features. Models trained on Creamsfields are tested on VFestival and vice-versa. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 11 (McNemar's test with 95% of confidence interval).

Train/Test	**Creamfields/VFestival**			**VFestival/Creamfields**		
Task	$\text{Model}_{\text{aggv,nfv}^{(\text{top})}}$	Accuracy		$\text{Model}_{\text{aggv,nfv}^{(\text{top})}}$	Accuracy	
Before	$\text{LR}^{\text{both}}_{\text{max,sum}^{(3)}}$	**0.800**	(+0.4%)	$\text{LR}^{\text{both}}_{\text{max,sum}^{(3)}}$	**0.872***	(+0.7%)
	$\text{GBDT}^{\text{w2v}}_{\text{mix,mean}^{(3)}}$	0.793	(+0.4%)	$\text{GBDT}^{\text{w2v}}_{\text{sum,sum}^{(1)}}$	0.861	(+2.6%)
During	$\text{LR}^{\text{both}}_{\text{max,max}^{(1)}}$	0.707	(+0.5%)	$\text{LR}^{\text{both}}_{\text{sum,max}^{(1)}}$	0.757*	(+1.6%)
	$\text{GBDT}^{\text{w2v}}_{\text{max,max}^{(1)}}$	**0.746***	(+2.2%)	$\text{GBDT}^{\text{both}}_{\text{mean,mix}^{(1)}}$	**0.778***	(+3.5%)
After	$\text{LR}^{\text{both}}_{\text{mix,max}^{(1)}}$	**0.811**	(+2.4%)	$\text{LR}^{\text{w2v}}_{\text{mean,sum}^{(3)}}$	0.811*	(+2.4%)
	$\text{GBDT}^{\text{both}}_{\text{sum,sum}^{(5)}}$	**0.811***	(+2.2%)	$\text{GBDT}^{\text{both}}_{\text{sum,mean}^{(5)}}$	**0.817***	(+1.0%)

5.2.3.2 Contribution of word embedding features

The experiments above show that the robustness of our classifiers across events is enhanced when word embedding features capturing text semantics for positive and negative attendance are introduced.

To analyze this effect, we consider the twenty five most important terms (BoW features) occurring in the Creamsfields and VFestival datasets and used by the GBDT classifiers trained on the corresponding dataset for each one of the three tasks. Term importance is determined by the gain in the loss function when the node of a decision tree is split on that feature (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Then, for each task, the terms occurring in both the datasets are filtered out since they are non-relevant for our analysis. Finally, the Cosine similarity between the embedding vectors of each pair in the Cartesian product of the remaining terms is computed.

The results of this investigation are summarized in Table 13, which reports the top-10 pairs of terms with the highest similarity. From the table, it can be seen that the two datasets include different terms that are likely to be relevant for the classification of the post and whose semantic is captured by the word embedding. For example, the word 'purchase', which appears in some posts of Creamfields but not in the VFestival dataset, has a similar embedding vector to the word 'sell' which, in turn, appears in the VFestival dataset but not in Creamfields: both words are mainly used in posts related to the purchase of the tickets for the events. For the *during* task, we can observe a high similarity between the embedding of the words 'excite' and 'amaze' and also 'excitement' and 'atmosphere': in both cases, the words mainly represent the attendees' experiences during the event. Similarly for the *after* task, where we can see the similarity between the words representing periods of time as 'week' and 'weekend' used mainly to refer to the past event.

In summary, in addressing RQ2, we find that the w2v and NFV features introduced al-

low us to exploit the semantic similarity of text, thus improving the classification accuracy and the robustness across events of our classifiers.

Table 13 – Per task top-10 most similar pairs of terms (according to the w2v vectors) in the sets of disjoint terms occurring in the Creamsfields and VFestival datasets.

Before			**During**			**After**		
Creamsfields	VFestival	Sim.	Creamsfields	VFestival	Sim.	Creamsfields	VFestival	Sim.
purchase	sell	0.656	excite	amaze	0.545	week	weekend	0.713
want	go	0.452	wait	watch	0.432	week	day	0.655
ready	wait	0.432	back	rest	0.410	hear	listen	0.649
camp	tent	0.431	excitement	atmosphere	0.382	ago	years	0.505
want	wait	0.419	go	rest	0.347	leeds	justin	0.453
ready	unprepared	0.402	jealous	sick	0.345	week	years	0.433
dj	buzzin	0.391	excitement	experience	0.341	week	time	0.408
work	go	0.354	go	jump	0.318	ago	old	0.393
want	bring	0.315	buzz	atmosphere	0.317	go	miss	0.389
ready	finally	0.300	go	watch	0.317	good	little	0.389

5.2.4 Results: RQ3

Our last research question (RQ3) asks if it is possible to identify expressions commonly used by users on social media to express attendance (or not) to an event. By using our whole corpus of gathered tweets, we conduct a co-occurrence analysis of the words written in the user's posts. First, by using our most accurate classifiers for each event and task, we classify all of the unlabeled tweets into (a) attendance and (b) not attendance. Then, for each class, task and event, we compute term co-occurrences to find the set of words most frequently co-occurring in posts of the same class, task and event. The results of this analysis are shown in Table 14 for the positive attendance class, and in Table 15 for the negative one. For the sake of simplicity, in both tables we report only the top-5 sets of 3 words ordered by their co-occurrence frequency. Note that to compute the co-occurrence frequencies, we do not consider the order in which the words occur. It is also worth noting that in this analysis all of the numeric values have been replaced with the symbol '#'.

Looking at Table 14, for the *before* task, we clearly notice the user's expectation to attend the event when they count down the days, reflected by a high occurrence of the set "{#, days, until}", or when they mention future participation, supported by the high frequency of the set "{be, next, week}". This is illustrated for example in the following posts found in the Creamfields dataset: (a) *"I'll be at Creamfields this time next week and I cannot wait"*; (b) *"This time next week I'll be in Creamfields, what an absolute blinding feeling"*; (c) *"Can't believe Creamfields is next week"*. For the *during* task, the co-occurrence of the words has a much lower frequency. This is justified by the slightly lower amount of tweets in this temporal slot and also by the higher diversity of manners in which people express their current attendance: for example, they sometimes post photos

with very few words to describe their personal experience. For the *after* task, we can see a similar style of posts for both events, expressing pleasure and happiness for attending the event: *"weekend, best, had"*, and desires to relive such experience, commonly written by using the expression *"take me back"*.

On the other hand, the sets of words reported in Table 15 help us to devise common expressions for the negative attendance case. In particular, for the *before* task and in both datasets, we notice a high correlation among the words 'pounds' and 'ticket' associated with the ticket cost and time periods like month, weekend or day. Indeed, these words are mostly used in advertisements tweets of sponsors and ticket sellers, which are not considered to be actual attendees. For the *during* task, we notice in the Creamfields dataset common expressions of people regretting not being able to attend the event: (a) *"Couldn't be anymore gutted that I'm not going to Creamfields, cry cry cry"*; (b) *"gutted not to be back at Creamfields this year"*; (c) *"A part of me is very gutted not to be heading to Creamfields tomorrow"*. For the *during* and *after* tasks, we observe that many non-attendance posts contain terms related to the performance of famous artists. Those posts are, in general, written by sponsors, newspapers and fans not necessarily attending the festival.

Table 14 – Top-5 most frequent 3-grams in the positive attendance class.

	Creamfields		**VFestival**	
Task	Words	Freq.	Words	Freq.
Before	{be, next, week}	253	{#, days, until}	414
	{next, week, time}	214	{#, days, till}	59
	{be, next, time}	178	{ be, so, excited}	59
	{be, week, time}	173	{#, only, hours}	50
	{#, days, work}	170	{weekend, so, excited}	44
During	{#, more, sleep}	68	{park, chelmsford, highlands}	45
	{up, line, great}	40	{you, so, proud}	21
	{#, uk, kingdom}	31	{you, much, thank}	14
	{#, uk, united}	31	{so, park, hylands}	14
	{we, here, come}	29	{down, via, chilling}	13
After	{my, best, life}	377	{weekend, best, had}	302
	{me, back, take}	317	{me, back, take}	207
	{my, weekend, best}	312	{my, weekend, best}	174
	{last, time, week}	283	{my, best, life}	147
	{was, last, time}	233	{weekend, good, such}	134

Table 15 – Top-5 most frequent 3-grams in the negative attendance class.

	Creamfields		**VFestival**	
Task	Words	Freq.	Words	Freq.
Before	{#, day, pounds}	7894	{#, tickets, pounds}	141
	{#, pounds, monthdate}	3328	{#, weekend, pounds}	140
	{#, pounds, warrington}	3316	{#, ticket, pounds}	127
	{#, monthdate, warrington}	3316	{#, pounds, sale}	118
	{#, day, camping}	3122	{#, camping, pounds}	118
During	{festival, man, dies}	345	{justin, is, performing}	174
	{be, not, going}	283	{great, john, newman}	120
	{be, not, gutted}	234	{justin, bieber, staffordshire}	116
	{festival, music, dance}	223	{justin, not, bieber}	112
	{going, was, wish}	196	{you, so, love}	101
After	{#, mix, essential}	288	{justin, monthdate, performing}	400
	{#, cirez, essential}	233	{justin, performing, staffordshire}	271
	{#, cirez, mix}	232	{justin, united, kingdom}	262
	{cirez, mix, essential}	209	{justin, monthdate, staffordshire}	260
	{festival, man, dies}	177	{justin, performing, united}	259

5.3 Example Application: Transport Planning

As an example use case for our proposed classifiers, we aim to evaluate the geographic areas with a higher potential demand for transportation services to an event. We analyse the hometown of users who have been predicted to attend a given festival by our classifiers. This analysis can be useful to support strategies for the allocation of shuttle buses or ride-sharing services to the event, or to forecast possible traffic congestions towards the event. We conduct this analysis upon our Creamfields dataset, the largest in terms of users, thereby allowing for a more realistic analysis compared to the VFestival dataset.

Starting from the event-related posts, we aim to infer the users who participated in the festival. For this purpose, it is important to note that often users on social media share more than one post related to a given event. Each post can be classified as attendance or not attendance depending on the content. There is no guarantee that all event-related posts of the same user will be consistently classified as attending or not attending. We therefore need to infer, given a number of posts of the same user, possibly not uniformly classified as attendance or not attendance, if the user is actually attending or not the event.

For the purpose of this example application, we trained our attendance classifiers on the Creamfields labeled data. We applied the best model for each task according to Table 9 to classify the whole dataset of about 90k tweets. We were able to predict as positive a total of 35,239 tweets. Distinguishing users attending or not attending from a number of - possible discordant - posts can be done in several ways, for example, through majority voting. We propose here a slightly more sophisticated method taking into account the confidence of the used classifier in labelling each post. Intuitively, a more confident

attendance prediction should count more than a less confident one. Therefore, during the classification process, for each post, we keep the difference of the confidence scores between the attendance and non-attendance classes. Notice that this value ranges from -1 to 1, where 1 means a higher confidence score for the attendance class, and -1 means the lowest attendance score. Then, taking all the classified posts or users, we compute the mean of the difference of the confidence scores. Our intuition is to capture the most discordant users regarding attendance. As a final decision, we have two cases: (a) users with a positive mean have attended the event (b) users with zero or negative mean have not attended the event.

We perform two kinds of analysis. The first analysis is aimed at inferring future participation in the event based on the posts shared **b**efore the event. In the second analysis we also consider the posts shared during and after the event. The idea here is to use historical data to identify cities with high amount of attendees to support future strategies in transportation and advertisement for the next editions of the event. The first analysis is based only on the posts published before the event. The idea here is to predict which are the geographical areas with the highest quantity of attendees who may potentially be needing transportation services to reach the event location. We recall that Creamfields is held in Daresbury, England, located between Liverpool and Manchester. We apply the above approach considering only the posts published at least one day before the event. From the quantity of inferred attendees, we collected, using the Twitter REST API, a total of 3856 users' profiles containing details of the users' hometown within their Twitter profiles. Figure 17 shows the spatial distribution of the inferred attendees of the Creamfields festival.

As expected, the results indicate a highest amount of participants in the surroundings of the event location as in the cities of Manchester and Liverpool. However, we can also identify other considerable amount of predicted attendees located in further cities such as London, Newcastle, Peterborough, Glasgow and Edinburgh. Intuitively, the higher the quantity of attendees, the higher the potential demand for transportation services in that area. Therefore, such information could be useful for generating an optimized planning of bus routes across cities and this can provide efficient transportation services to the event. Ride-sharing applications could also take advantage from the identification of groups of predicted attendees. However, we leave such applications as possible future work.

For our second analysis, we run the best classifiers obtained in the generalization experiment for each of the three tasks on the relative sets of posts, namely before, during and after Our intention here is to identify cities with high amount of attendees to support future transportation and marketing strategies for the next editions of the event. Here, we use the approach described above to label a user as attendee or not, based on his/her posts. Table 16 summarizes the amount of inferred attendees by city. We have identified a total of 10788 inferred participants to the event that have also their hometown infor-

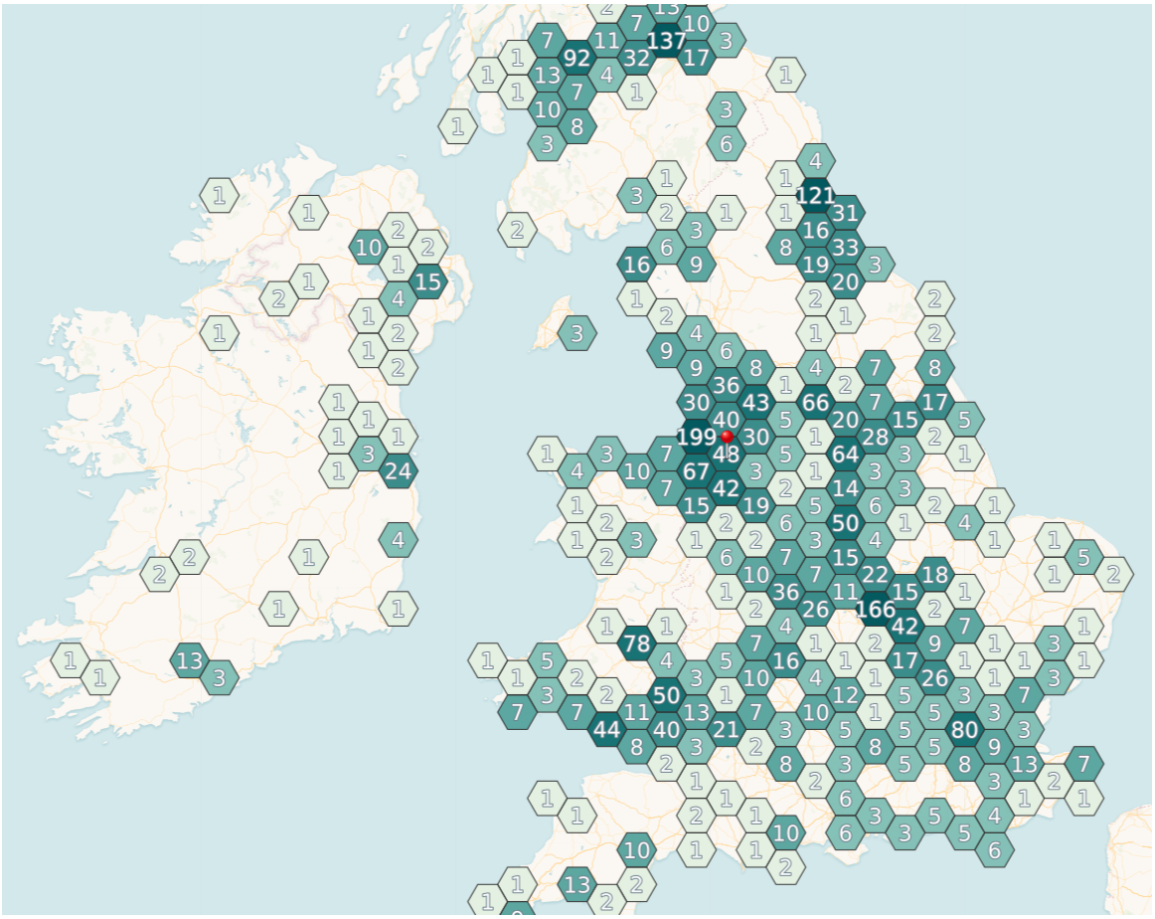


Figure 17 – Spatial distribution of inferred participants to the Creamfields festival (red point) from posts published before the event.

Table 16 – Distribution of the before, during and after inferred attendees at the Creamfields festival by hometown.

City	# attendees	City	# attendees	City	# attendees
Aberdeen	57	**Edinburgh**	**263**	Northampton	53
Birmingham	96	**Glasgow**	**221**	Nottingham	112
Bristol	114	Hull	62	Plymouth	86
Cambridge	53	Leeds	160	Sheffield	125
Cardiff	85	Leicester	88	South Wales	109
Coventry	67	**Liverpool**	**732**	Sunderland	55
Derby	53	**London**	**456**	Swansea	106
Doncaster	83	**Newcastle**	**312**	Warrington	150

mation displayed on their public Twitter profiles. Through the results, we can observe that the previous analysis, predicting the most transportation demanding areas, approximates well the final distribution of attendees by city. We note that Liverpool, Manchester and the surrounding area of the "North of England" present a high number of attendees. The Scottish cites of Edinburgh and Glasgow might require long-distance transportation services due to the distance of these cities to the event's location.

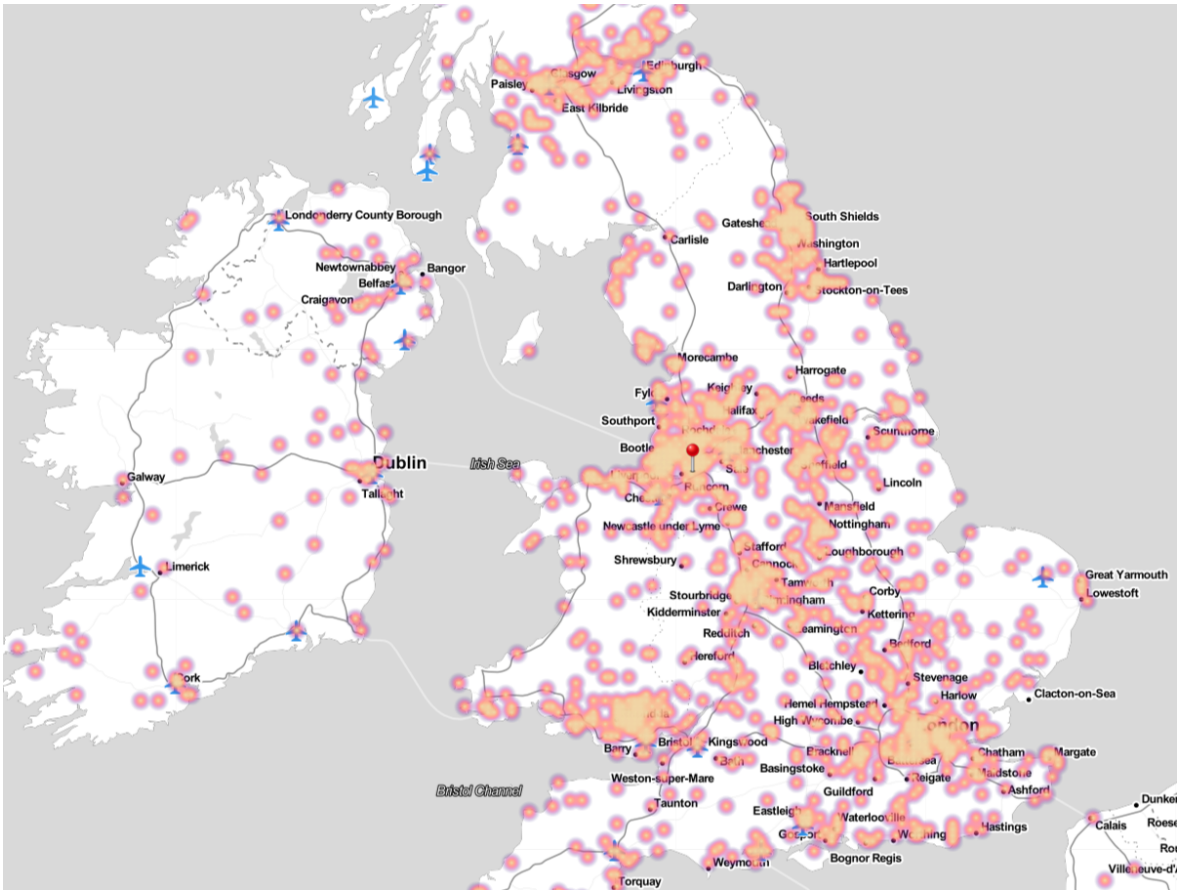


Figure 18 – Heatmap with distribution by hometown of the inferred attendees at the Creamfields festival (red point).

We provide a visualization of the results on a heat map in Figure 18. We also visualized the airports that connect cities from Ireland and The Netherlands to the UK, in blue. Looking at this visualization, we observe red areas (i.e. the hot regions) with a higher density of hometowns of the inferred festival attendees. We observe that, as expected, most of the dense areas are close to the event location. However, we also note some small dense areas located in cities outside the UK, such as the Irish cities of Dublin, Cork and Belfast and the Amsterdam and The Hague Dutch cities. The attendees from these areas might first fly to airports in the UK.

5.4 Final Considerations

In this chapter, we proposed a classification approach to infer event attendance from users media posts with the final objective of estimating transportation demand. A key detail of our proposed approach is that our inference is done by classifying the non-geotagged content of the users' posts. By not relying on geotagged posts we can analyze a much larger number of posts to predict user attendance to a given event. The large basis of users covered by our approach makes it a great candidate to enable innovative services

and applications in the field, for example, of transportation planning and crowd safety management. We structured the attendance inference into three distinct classification tasks to identify the attendance from the posts published before, during and after the event.

We trained machine-learning classifiers using tweets related to two large music festivals in the UK, and we evaluated their accuracy, precision and recall. The results discussed in Section 5.2.2 show that our approach provides remarkably good performance, exhibiting $\sim$91% accuracy at classifying users that have indicated their intention to attend the event. Our analysis showed that word embedding features contribute importantly to the performance. Additionally, we highlighted the most informative group of features and assessed the accuracy of our classifier even on an objective test set constituted by geo-tagged tweets. In Section 5.2.3, we analyzed the generalization of the learned models across the datasets and propose additional word embedding features to improve cross-dataset performance. Furthermore, in Section 5.2.4 we investigated the common expressions used by social media users to express (or not) attendance to an event. Finally, in Section 5.3, we proposed an example of application of our methodology in event-related mobility demand. The application derives the rides demand for the Creamfields festival based on the inferred attendees and exhibits the number of potential demand by cities.

6 Conclusions

Ride sharing consists in the sharing of a vehicle by two (or more) persons who move along similar itineraries and time schedules. The prearrangement process to match the supply and demand is a key characteristic of ride sharing. Many works in literature have proposed ride sharing solutions to avoid single occupancy vehicle trips. In this context, there are two main types of services for ride sharing: *matching agencies* and *transportation service operators*. The *matching agencies* exploit different kinds of *matching algorithms* to find ride matchings between individual car drivers and passengers,while *transportation service operators* provide ride sharing services identifying and supplying the demand of rides with their own vehicles and drivers, such as airport shuttles. This thesis has research contributions in both these scenarios: we propose a ride matching algorithm for alternative destinations and propose a classification method to identify potential users for a transportation service towards large events.

Along the first direction, Chapter 4 introduced the Activity-Based Ride Matching algorithm (ABRM). The ABRM has shown optimistic results to improve the efficacy of traditional ride sharing systems. Most of the ride matching algorithms in literature are typically based on the spatial and temporal aspects of the rides. However, the key idea of the ABRM is to consider also rides to alternative destinations based on the intended activity of the passenger. The ABRM is motivated by recent studies about human mobility that highlight the individual tendency of the people to be regular or not in choosing the places where to perform some activities. Thus, the investigations presented in this chapter focus in the following research question: *"Can the usage of ride sharing systems be boosted by exploiting alternative destinations based on the intended activity of the passenger?"*.

The contributions in the second direction are presented in Chapter 5, where we introduced an approach to exploit the content of non-geotagged posts on social media to infer the user attendance to large events. This approach is motivated by the fact that large events cause the movement of thousands of people to a specific location, requiring a proper transportation plan to supply the demand of rides. However for large events such as music concerts and football matches, the identification of ride demands might not be a straightforward process. The key research question of this chapter is: *"Is it possible to infer past, current and future user attendance to large events through posts on social media to forecast the demand of rides?"*.

The remaining sections of this chapter are organized as follows. Section 6.1 summarizes the main contributions of this thesis, Section 6.2 discusses some research limitation of the studies made in this work and Section 6.3 gives possible directions for future research. Finally, Section 6.4 presents the list of publications produced during this PhD programme.

6.1 Thesis Contributions

In the following we discuss the contributions of this thesis:

Improving the efficacy of ride matching algorithms. In Chapter 4 we proposed the ABRM algorithm which is aimed at matching users' ride requests with ride offers reaching alternative destinations where the intended user activity can be performed. Most of the existing methods, although very sophisticated, suffer from limitations in the use of semantic information, such as the passenger's intended activity. Therefore, our approach can be seen as complementary to the existing methods, providing a different perspective. Here, the intended activity is considered as the target of the ride sharing instead of a fixed destination. Our assumption is that having alternative destination increases the ride sharing possibilities. We investigated this assumption in our experiments. We exploited two large datasets of Foursquare containing check-ins performed in the cities of New York and Tokyo and built two semi-synthetic datasets representing disjoint sets of *Ride Requests* and *Ride Offers* called: NYC and TKY. In Section 4.4.3, the experiments conducted on these datasets showed that ABRM can boost in average up to 54.69% the efficacy on finding compatible ride matching with respect to traditional fixed destination-oriented ride sharing. Section 4.2.1 proposed four features to rank the matchings returned by ABRM, namely: time delay, distance to walk, ride duration and ride length. With these features we intend to model how much the rides to alternative destinations can meet the requirements specified in the user request. Section 4.4.4 discusses the contribution of these ride features for the evaluation of the qualities of the rides retrieved by the proposed matching algorithm. Specifically, for the TKY dataset and the time delay feature, the results showed that even when we consider matches with only 10 minutes delay, more than half of the requests could potentially be satisfied toward an alternative destination. Another interesting analysis discussed in Section 4.4.6 shows the activities that most favor the ride sharing in the alternative destination scenario. We observed, for NYC, a boost on ride sharing possibilities for activities related to "Italian Restaurant" and "Bar" with an improvement of +58.02% and +57.86% respectively. In turn, for TKY, activity "Food & Drink Shop" achieves the highest boost with +64.28%. This insight confirm the intuition that entertainment or eating are in general the activities that can most benefit from the proposed approach. In the last part of the chapter, we introduce ComeWithMe which is a carpooling system offering alternative destinations designed upon the ABRM algorithm.

Identifying transportation demands for large events. In Chapter 5 we proposed a classification approach to infer attendance to events from the users' media posts. A key detail of our proposal is that our inference is done by classifying the non-geotagged content of the users' posts. Thus, we can analyze a much larger number of posts to predict user attendance to a given event compared to the sparse geotagged posts. The large base of users covered by our approach enables innovative services and applications like transportation planning and crowd safety management. We structured the attendance

inference into three distinct classification tasks to identify the attendance from the posts published before, during and after the event. We trained machine-learned classifiers using tweets related to two large music festivals in the UK, and we evaluated their accuracy, precision and recall. The results discussed in Section 5.2.2 show that our approach provides a remarkably good performance, exhibiting ∼91% accuracy at classifying users who have indicated their intention to attend the event. Our analysis showed that *word embedding* features contribute saliently to the performance of our classifiers. Additionally, we highlighted the most informative group of features and assessed the accuracy of our classifier even on an objective test set composed of geo-tagged tweets. In Section 5.2.3, we analyzed the generalization of the learned models across the datasets and proposed additional word embedding features to improve cross-dataset performances. For example, when classifying the posts published after the event, by including both the *word embedding* and *Normalized Frequency Vectors (NFV)* features, the Gradient Boosting Decision Trees (GDBT) algorithm has increased up to +7.8% (from 73.3% to 81.1%) its generalization ability when trained on Creamfields dataset and tested on VFestival dataset. Furthermore, in Section 5.2.4, we investigated the common expressions used by social media users to express (or not) attendance to an event. Finally, in Section 5.3, we presented an example of application of our methodology in event-related transportation. This proposed application aimed to evaluate the geographic areas with a higher potential demand for transportation services to an event. We analyse the hometown of users who have been predicted to attend a given festival by our classifiers. This analysis can be useful to support strategies for the allocation of shuttle buses or ride sharing services to the event, or to forecast possible traffic congestion towards the event.

The next section discusses some limitation of these studies.

6.2 Research Limitations

In the following, we discuss some limitations of this thesis:

As shown in Chapter 4, the Activity-Based Ride Matching algorithm (ABRM) has boosted the amount of ride matchings by considering also rides to alternative destinations to supply the ride requests. However, some limitations may exist in this study. Indeed, the analysis exhibited in Chapter 4 were limited to study the improvement in efficacy of the proposed algorithm, we did not evaluate the efficiency of the algorithm. Another point that requires further investigation is that our analysis is based on the assumption that users might be flexible in their mobility habits and provided interesting insights on the extent to which a ride sharing service could take advantage of this spontaneous attitude. We notice, however, that an ad-hoc study with real users would be necessary to have a reliable measure of the actual acceptance of ride offers towards alternative destinations. Another research limitation of our work it that we do not deal with the routing allocation

problem neither consider the number of seats available in the cars offering the rides.

In turn, the Chapter 5 introduced a machine learning approach for identifying transportation demands for large events. The proposed approach exploits the content of non-geotagged posts on social media to infer the attendance of large events. Although the approach is suitable to different kind of events (e.g. sportive event, music festival, scientific conferences, etc) we have assessed its prediction performance conducting experiments in only one context of event: music festivals. Therefore, a further investigation would be necessary mainly to understand how effective are the embedding features to improve accuracy when classifying event attendance in different context of events and how good the models can generalize. Therefore, further study would be necessary to validate the performance of our approach applied to different kind of events.

The next section discusses some possible future work of the contributions of this thesis.

6.3 Future Work

The studies performed in this thesis open up a wide spectrum of improvement opportunities. We thus discuss possible directions for future research according to the main contributions of this thesis:

Improving the efficacy of ride matching algorithms. The ABRM algorithm largely increases the number of candidate rides by considering a set of alternative destinations. The ABRM algorithm opens space for research on efficient approaches for ride matchings retrieval considering now the new proposed semantic dimension: the intended user's activity. The ABRM algorithm could be extended to consider also the returning of the passenger to the pick-up point. For example, frequently occurs that a person can move from her home to a destination, for instance, a supermarket, and she may want to return back home after perform her intended activity. An interesting point to be studied as future work is to refine the evaluation method taking into account the individual willingness to change the destination to take a ride. In fact, the suggestion for the alternative destination could be refined considering user's preferences. We would like to use a ranking-based solution which orders the potential ride-offers on the basis of a weighted combination of a set of features modeling the different aspects of user satisfaction. An advantage of this solution is that we can investigate more in depth the effect of user flexibility in accepting changes involving the following dimensions: the desired departure time, the distance from the pick up point and, as novelty, the recommendations of the alternative destinations. Thus, the idea here is that the ride matching with higher ranking scores should attain higher user satisfaction. For this purpose, we also aim at building machine-learned ranking (MLR) models which is the application of machine learning, typically supervised, in the construction of ranking models for information retrieval systems. In this case, the training data consists of ride requests and ride offers matching them together with relevance

degree of each match. Furthermore, some opened question concerning data privacy must be addressed to apply the ABRM in current ride-sharing systems, such as Lyft and Uber. For the ABRM algorithm to achieve a more refined ranking model, it would be valuable to collect users' preferences with respect to the places that the users visited to perform specific activities. However, this is sensitive information requiring a careful review of data privacy.

Identifying transportation demands for large events. As future work, we aim to improve the results of our classifiers using information extracted from the visual content of the published photos and videos. In fact, our proposed approach has explored mainly the textual features through the use of embedding to derive new useful features to the classifiers. However, further investigation of how media content could be further explored to extract relevant features for inferring attendance are still necessary. The analysis of visual content is a growing trend in social media and could be better explored in our classification process through the use of deep learning techniques. We also plan to use deep learninig techniques for our proposed classification task. The idea is to use a bigger amount of data, collected from different kinds of events, to enable a comparison between shallow models and deep models. Another open research question to be investigated is: "how to use learned models to predict attendance to other events, especially not similar events"? One path could be to explorer transfer learning techniques for using knowledge obtained while solving one problem and applying it to a different but related problem. For example, the knowledge gained while learning how to predict attendees for music festival could be useful when classifying attendees for sportive events. A similar situation occurs when classifying posts containing textual content from different languages. For example, when using the knowledge learned from event-related posts written in English to predict attendance of posts written in Spanish. Furthermore, we aim to further explore our methodology in the context of ride sharing. In this context, a plausible research question could be: "how can we take advantage of predicting the attendees of a event to propose groups of carpooling with optimal affinity between the individuals"? Here, the affinity represents the interesting in common topics. Thus, one idea is to use our classifiers to recommend individuals to join their trips to the event matching users with similar social media preferences. The objective of grouping the user could represent the optimization of two simultaneous function: minimize the total number of vehicle used to transport the attendees to the event and maximize the enjoyability experienced by the users traveling together.

The next section lists the scientific publications achieved.

6.4 List of publications

In this section, we list all the references to published papers produced in the course of the Ph.D. programme. The list of the papers is organized according to the two main research objectives of this thesis.

1) Improving the efficacy of ride matching algorithms.

- **Boosting Ride Sharing With Alternative Destinations.** Vinicius Monteiro de Lira, Raffaele Perego, Chiara Renso, Salvatore Rinzivillo, Valéria Cesário Times. *IEEE Trans. Intelligent Transportation Systems, Volume 19. pp. 2290-2300 (2018).*
- **The ComeWithMe System for Searching and Ranking Activity-Based Carpooling Rides.** Vinicius Monteiro de Lira, Chiara Renso, Raffaele Perego, Salvatore Rinzivillo, Valéria Cesário Times. *In Proceedings of the 39th International ACM conference on Research and Development in Information Retrieval, pp. 1145-1148.* (ACM SIGIR) Pisa, Italy July 17 - July 21, 2016.
- **Searching and Ranking Activity-based Carpooling Ride** Chiara Renso, Salvatore Rinzivillo, Valéria Cesário Times, Vinicius Monteiro de Lira, Raffaele Perego. *Extended abstract in 7th Italian Information Retrieval Workshop (IIR),* Venice, Italy May 20 - May 31, 2016.
- **Activity-based Carpooling with ComeWithMe.** Vinicius Monteiro de Lira, Salvatore Rinzivillo, Valéria Cesário Times, Chiara Renso, Raffaele Perego. *24th Italian Symposium on Advanced Database Systems, pp. 142-149,* (SEBD) Ugento, Lecce, Italy, June 19-22, 2016.
- **ComeWithMe: An activity-oriented carpooling approach**. Vinícius Monteiro de Lira, Valeria Cesario Times, Chiara Renso, Salvatore Rinzivillo. *IEEE 18th International Conference on Intelligent Transportation Systems, pp. 2574-2579,* ITSC 2015.

2) Identifying transportation demands for large events.

- **Event attendance classification in social media.** Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times. *Information Processing & Management Journal, IPM,* Elsevier, v. 56, n. 3, p. 687–703, 2019.
- **Exploring Social Media for Event Attendance.** Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times.

Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 447-450, ASONAM 2017, Sydney, Australia, July 31 - August 03, 2017.

References

AGATZ, N.; ERERA, A.; SAVELSBERGH, M.; WANG, X. Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research*, v. 223, n. 2, p. 295 – 303, 2012. ISSN 0377-2217. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0377221712003864>.

AGATZ, N. A.; ERERA, A. L.; SAVELSBERGH, M. W.; WANG, X. Dynamic ride-sharing: A simulation study in metro atlanta. *Transportation Research Part B: Methodological*, v. 45, n. 9, p. 1450 – 1464, 2011. ISSN 0191-2615. Select Papers from the 19th ISTTT. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0191261511000671>.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: _____. *Mining Text Data*. [S.l.]: Springer, 2012. p. 163–222.

ALBUQUERQUE, F. C.; CASANOVA, M. A.; LOPES, H.; REDLICH, L. R.; MACEDO, J. A. F. de; LEMOS, M.; CARVALHO, M. T. M. de; RENSO, C. A methodology for traffic-related twitter messages interpretation. *Computers in Industry*, Elsevier, v. 78, p. 57–69, 2016.

ALKANDARI, A.; ALNASHEET, M.; ALSHEKHLY, I. F. T. Smart cities: Survey. v. 2, 01 2012.

ALONSO-MORA, J.; SAMARANAYAKE, S.; WALLAR, A.; FRAZZOLI, E.; RUS, D. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 114, n. 3, p. 462–467, 2017. ISSN 0027-8424. Disponível em: <http://www.pnas.org/content/114/3/462>.

BAKERMAN, J.; PAZDERNIK, K.; WILSON, A.; FAIRCHILD, G.; BAHRAN, R. Twitter geolocation: A hybrid approach. *ACM TKDD*, v. 20, n. 3, p. 34:1–34:17, 2018.

BALIKAS, G.; AMINI, M.-R. An empirical study on large scale text classification with skip-gram embeddings. *arXiv preprint arXiv:1606.06623*, 2016.

BENEVOLO, C.; DAMERI, R. P.; D'AURIA, B. Smart mobility in smart city. In: _____. *Empowering Organizations: Enabling Platforms and Artefacts*. Cham: Springer International Publishing, 2016. p. 13–28. ISBN 978-3-319-23784-8. Disponível em: <https://doi.org/10.1007/978-3-319-23784-8_2>.

BERGSTRA, J.; BARDENET, R.; BENGIO, Y.; KÉGL, B. Algorithms for hyper-parameter optimization. In: *Proc. of NIPS*. [S.l.: s.n.], 2011. p. 2546–2554. ISBN 978-1-61839-599-3.

BERLINGERIO, M.; GHADDAR, B.; GUIDOTTI, R.; PASCALE, A.; SASSI, A. The graal of carpooling: Green and social optimization from crowd-sourced data. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 80, p. 20–36, 2017.

BOTTA, F.; MOAT, H. S.; PREIS, T. Quantifying crowd size with mobile phone and twitter data. *Royal Society open science*, The Royal Society, v. 2, n. 5, p. 150–162, 2015.

CAMPANA, M. G.; DELMASTRO, F.; BRUNO, R. A machine-learned ranking algorithm for dynamic and personalised car pooling services. In: *IEEE Intelligent Transportation Systems Conference*. [S.l.: s.n.], 2016.

CARPINETO, C.; ROMANO, G. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, v. 44, n. 1, p. 1:1–1:50, 2012.

CAULFIELD, B. Estimating the environmental benefits of ride-sharing: A case study of dublin. *Transportation Research Part D: Transport and Environment*, v. 14, n. 7, p. 527 – 531, 2009. ISSN 1361-9209. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1361920909000893>.

CESARIO, E.; CONGEDO, C.; MAROZZO, F.; RIOTTA, G.; SPADA, A.; TALIA, D.; TRUNFIO, P.; TURRI, C. Following soccer fans from geotagged tweets at FIFA World Cup 2014. In: *Proc. of IEEE ICSDM 2015*. [S.l.: s.n.], 2015.

CESARIO, E.; IANNAZZO, A. R.; MAROZZO, F.; MORELLO, F.; RIOTTA, G.; SPADA, A.; TALIA, D.; TRUNFIO, P. Analyzing social media data to discover mobility patterns at EXPO 2015: Methodology and results. In: *Proc. of HPCS 2016*. [S.l.: s.n.], 2016.

CESARIO, E.; MAROZZO, F.; TALIA, D.; TRUNFIO, P. Sma4td: A social media analysis methodology for trajectory discovery in large-scale events. *Online Social Networks and Media*, Elsevier, v. 3-4, p. 49 – 62, 2017.

CHAN, N. D.; SHAHEEN, S. Ridesharing in north america: Past, present, and future. v. 32, p. 93–112, 01 2012.

CHANG, H.-w.; LEE, D.; ELTAHER, M.; LEE, J. @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In: *Proc. of IEEE/ACM ASONAM 2012*. [S.l.: s.n.], 2012.

CHEN, C.; MA, J.; SUSILO, Y.; LIU, Y.; WANG, M. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, Elsevier, v. 68, p. 285 – 299, 2016.

CHENG, Z.; CAVERLEE, J.; LEE, K. You are where you tweet: A content-based approach to geo-locating twitter users. In: *Proc of ACM CIKM 2010*. [S.l.: s.n.], 2010.

CHIANG, M.-F.; LIM, E.-P.; LEE, W.-C.; HOANG, T.-A. Inferring trip occupancies in the rise of ride-hailing services. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2018. (CIKM '18), p. 2097–2105. ISBN 978-1-4503-6014-2. Disponível em: <http://doi.acm.org/10.1145/3269206.3272025>.

CHO, E.; MYERS, S. A.; LESKOVEC, J. Friendship and mobility: User movement in location-based social networks. In: *Proc. of ACM SIGKDD 2011*. [S.l.: s.n.], 2011.

CHO, S.; KANG, J.-Y.; YASAR, A.-U.-H.; KNAPEN, L.; BELLEMANS, T.; JANSSENS, D.; WETS, G.; HWANG, C.-S. An Activity-based Carpooling Microsimulation Using Ontology. *Procedia Computer Science*, Elsevier B.V., v. 19, n. Ant, p. 48–55, jan. 2013. ISSN 18770509. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1877050913006200>.

CICI, B.; MARKOPOULOU, A.; FRIAS-MARTINEZ, E.; LAOUTARIS, N. Assessing the potential of ride-sharing using mobile and social data: A tale of four cities. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. New York, NY, USA: ACM, 2014. (UbiComp '14), p. 201–211. ISBN 978-1-4503-2968-2. Disponível em: <http://doi.acm.org/10.1145/2632048.2632055>.

CORREIA, G.; VIEGAS, J. M. Carpooling and carpool clubs: Clarifying concepts and assessing value enhancement possibilities through a stated preference web survey in lisbon, portugal. *Transportation Research Part A: Policy and Practice*, v. 45, n. 2, p. 81 – 90, 2011. ISSN 0965-8564. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0965856410001552>.

DAILEY, D.; LOSEFF, D.; MEYERS, D. Seattle smart traveler: dynamic ridematching on the world wide web. *Transportation Research Part C: Emerging Technologies*, v. 7, n. 1, p. 17 – 32, 1999. ISSN 0968-090X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0968090X99000078>.

D'ANDREA, E.; DUCANGE, P.; LAZZERINI, B.; MARCELLONI, F. Real-time detection of traffic from twitter stream analysis. *IEEE TITS*, v. 16, n. 4, p. 2269–2283, 2015.

DREWS, F.; LUXEN, D. Multi-hop ride sharing. In: *Sixth annual symposium on combinatorial search*. [S.l.: s.n.], 2013.

DU, R.; YU, Z.; MEI, T.; WANG, Z.; WANG, Z.; GUO, B. Predicting activity attendance in event-based social networks: Content, context and social influence. In: *Proc. of UbiComp 2014*. [S.l.: s.n.], 2016.

EFSTATHIADES, H.; ANTONIADES, D.; PALLIS, G.; DIKAIAKOS, M. D. Users key locations in online social networks: identification and applications. *Social Network Analysis and Mining*, Springer, v. 6, n. 1, p. 66, 2016.

EFTHYMIOU, D.; ANTONIOU, C. Use of social media for transport data collection. *Procedia - Social and Behavioral Sciences*, Elsevier, v. 48, p. 775 – 785, 2012.

FARKAS, K.; FEHER, G.; BENCZUR, A.; SIDLO, C. Crowdsending based public transport information service in smart cities. *IEEE Communications Magazine*, v. 53, n. 8, p. 158–165, August 2015. ISSN 0163-6804.

FEEHAN, P. Attendance at sports events. *Handbook on the economics of sport*, Edward Elgar Cheltenham, p. 90–99, 2006.

FLORIDA, R. *America's Ongoing Love Affair With the Car*. 2015. <https://www.citylab.com/transportation/2015/08/americas-continuing-love-affair-with-the-car/401474/>. Last access April 18, 2018.

FLÜGGE, B. Introduction. In: FLÜGGE, B. (Ed.). *Smart Mobility – Connecting Everyone: Trends, Concepts and Best Practices*. Wiesbaden: Springer Fachmedien Wiesbaden, 2017. p. 1–3. ISBN 978-3-658-15622-0. Disponível em: <https://doi.org/10.1007/978-3-658-15622-0_1>.

FURUHATA, M.; DESSOUKY, M.; ORDóñEZ, F.; BRUNET, M.-E.; WANG, X.; KOENIG, S. Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological*, v. 57, p. 28 – 46, 2013. ISSN 0191-2615. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0191261513001483>.

GAL-TZUR, A.; GRANT-MULLER, S. M.; KUFLIK, T.; MINKOV, E.; NOCERA, S.; SHOOR, I. The potential of social media in delivering transport policy goals. *Transport Policy*, Elsevier, v. 32, p. 115–123, 2014.

GAL-TZUR, A.; GRANT-MULLER, S. M.; MINKOV, E.; NOCERA, S. The impact of social media usage on transport policy: Issues, challenges and recommendations. *Procedia - Social and Behavioral Sciences*, v. 111, p. 937 – 946, 2014. ISSN 1877-0428. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877042814001293>.

GALIL, Z. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys (CSUR)*, ACM, v. 18, n. 1, p. 23–38, 1986.

GALLAND, S.; KNAPEN, L.; YASAR, A.-U.-H.; GAUD, N.; JANSSENS, D.; LAMOTTE, O.; KOUKAM, A.; WETS, G. Multi-agent simulation of individual mobility behavior in carpooling. *Transportation Research Part C: Emerging Technologies*, v. 45, p. 83 – 98, 2014. ISSN 0968-090X. Advances in Computing and Communications and their Impact on Transportation Science and Technologies. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0968090X14000035>.

GAO, L.; WU, J.; QIAO, Z.; ZHOU, C.; YANG, H.; HU, Y. Collaborative social group influence for event recommendation. In: *Proc. of ACM CIKM 2016*. [S.l.: s.n.], 2016.

GARGIULO, E.; GIANNANTONIO, R.; GUERCIO, E.; BOREAN, C.; ZENEZINI, G. Dynamic ride sharing service: are users ready to adopt it? *Procedia Manufacturing*, Elsevier, v. 3, p. 777–784, 2015.

GEISBERGER, R.; LUXEN, D.; NEUBAUER, S.; SANDERS, P.; VÖLKER, L. Fast detour computation for ride sharing. In: *ATMOS 2010 - 10th Workshop on Algorithmic Approaches for Transportation Modeling, Optimization, and Systems, Liverpool, United Kingdom, September 6-10, 2010*. [S.l.: s.n.], 2010. p. 88–99.

GEORGIEV, P.; NOULAS, A.; MASCOLO, C. The call of the crowd: Event participation in location-based social services. In: *Proc. of ICWSM 2014*. [S.l.: s.n.], 2014.

GUIDOTTI, R.; SASSI, A.; BERLINGERIO, M.; PASCALE, A.; GHADDAR, B. Social or green? a data-driven approach for more enjoyable carpooling. In: *IEEE 18th ITSC*. [S.l.: s.n.], 2015.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. 2. ed. [S.l.]: Springer, 2009.

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

HERBAWI, W.; WEBER, M. Modeling the multihop ridematching problem with time windows and solving it using genetic algorithms. In: *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*. [S.l.: s.n.], 2012. v. 1, p. 89–96. ISSN 1082-3409.

ISLAM, M.; AKTER, S.; IMRAN, M. M.; HOSSAIN, I.; HASSAN, N. Festival time transportation demand modeling in bangladesh. *Int. J. Sci. Eng. Investig*, v. 5, n. 48, p. 40–44, 2016.

JIANG, W.; DOMINGUEZ, C. R.; ZHANG, P.; SHEN, M.; ZHANG, L. Large-scale nationwide ridesharing system: A case study of chunyun. *International Journal of Transportation Science and Technology*, v. 7, n. 1, p. 45 – 59, 2018. ISSN 2046-0430. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2046043017300552>.

KAEWPITAKKUN, Y.; SHIRAI, K.; MOHD, M. Sentiment lexicon interpolation and polarity estimation of objective and out-of-vocabulary words to improve sentiment classification on microblogging. In: *Proceedings of the 28th Pacific Asia conference on language, information and computing*. [S.l.: s.n.], 2014.

KAISER, M. S.; LWIN, K. T.; MAHMUD, M.; HAJIALIZADEH, D.; CHAIPIMONPLIN, T.; SARHAN, A.; HOSSAIN, M. A. Advances in crowd analysis for urban applications through urban event detection. *IEEE TITS*, p. 1–21, 2017.

KELLEY, K. L. Casual Carpooling — Enhanced. *Journal of Public Transportation*, v. 10, n. 4, p. 119–130, 2007.

KINSELLA, S.; MURDOCK, V.; O'HARE, N. I'm eating a sandwich in Glasgow: Modeling locations with tweets. In: *Proc of SMUC 2011*. [S.l.: s.n.], 2011.

KITAMURA, R. An evaluation of activity-based travel analysis. *Transportation*, v. 15, n. 1, p. 9–34, Mar 1988. ISSN 1572-9435. Disponível em: <https://doi.org/10.1007/BF00167973>.

KORUPOLU, M. R.; PLAXTON, C. G.; RAJARAMAN, R. Analysis of a local search heuristic for facility location problems. *J. Algorithms*, v. 37, n. 1, p. 146–188, 2000. Disponível em: <https://doi.org/10.1006/jagm.2000.1100>.

LEE, K.; GANTI, R. K.; SRIVATSA, M.; LIU, L. When twitter meets foursquare: Tweet location prediction using foursquare. In: *Proc. of MOBIQUITOUS 2014*. [S.l.: s.n.], 2014.

LEETARU, K.; WANG, S.; CAO, G.; PADMANABHAN, A.; SHOOK, E. Mapping the global twitter heartbeat: The geography of twitter. *First Monday*, v. 18, n. 5, 2013.

LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, v. 3, p. 211–225, 2015.

LIN, J.; SASIDHARAN, S.; MA, S.; WOLFSON, O. A model of multimodal ridesharing and its analysis. In: *2016 17th IEEE International Conference on Mobile Data Management (MDM)*. [S.l.: s.n.], 2016. v. 1, p. 164–173.

LIRA, V. M. D.; TIMES, V. C.; RENSO, C.; RINZIVILLO, S. ComeWithMe: An activity-oriented carpooling approach. In: *IEEE 18th International Conference on Intelligent Transportation Systems*. [S.l.: s.n.], 2015.

LIRA, V. M. de; MACDONALD, C.; OUNIS, I.; PEREGO, R.; RENSO, C.; TIMES, V. C. Exploring social media for event attendance. In: *Proc. of IEEE/ACM ASONAM 2017*. [S.l.: s.n.], 2017.

LIRA, V. M. de; MACDONALD, C.; OUNIS, I.; PEREGO, R.; RENSO, C.; TIMES, V. C. Event attendance classification in social media. *Information Processing & Management*, Elsevier, v. 56, n. 3, p. 687–703, 2019.

LIRA, V. M. de; PEREGO, R.; RENSO, C.; RINZIVILLO, S.; TIMES, V. C. Boosting ride sharing with alternative destinations. *IEEE Trans. Intelligent Transportation Systems*, v. 19, n. 7, p. 2290–2300, 2018. Disponível em: <https://doi.org/10.1109/TITS.2018.2836395>.

LIRA, V. M. de; RINZIVILLO, S.; RENSO, C.; TIMES, V. C.; TEDESCO, P. C. A. R. Investigating semantic regularity of human mobility lifestyle. In: *IDEAS 2014, Portugal, July 7-9*. [S.l.: s.n.], 2014.

LIU, C. Y.; ZHOU, C.; WU, J.; XIE, H.; HU, Y.; GUO, L. Cpmf: A collective pairwise matrix factorization model for upcoming event recommendation. In: *Proc. of IJCNN 2017*. [S.l.: s.n.], 2017.

MA, S.; WOLFSON, O. Analysis and evaluation of the slugging form of ridesharing. In: *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 2013. (SIGSPATIAL'13), p. 64–73. ISBN 978-1-4503-2521-9. Disponível em: <http://doi.acm.org/10.1145/2525314.2525365>.

MACEDO, A. Q.; MARINHO, L. B.; SANTOS, R. L. Context-aware event recommendation in event-based social networks. In: *RecSys*. [S.l.: s.n.], 2015. ISBN 978-1-4503-3692-5.

MAHMUD, J.; NICHOLS, J.; DREWS, C. Home location identification of twitter users. *ACM TIST*, v. 5, n. 3, p. 47:1–47:21, 2014.

MARTIN, E.; SHAHEEN, S.; LIDICKER, J. Impact of carsharing on household vehicle holdings: Results from north american shared-use vehicle survey. *Transportation Research Record: Journal of the Transportation Research Board*, Transportation Research Board of the National Academies, n. 2143, p. 150–158, 2010.

MASOUD, N.; JAYAKRISHNAN, R. A real-time algorithm to solve the peer-to-peer ride-matching problem in a flexible ridesharing system. *Transportation Research Part B: Methodological*, v. 106, p. 218 – 236, 2017. ISSN 0191-2615. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0191261517301169>.

MASRI, A.; ZEITOUNI, K.; KEDAD, Z. Retry: Integrating ridesharing with existing trip planners. In: *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. New York, NY, USA: ACM, 2017. (SIGSPATIAL'17), p. 47:1–47:10. ISBN 978-1-4503-5490-5. Disponível em: <http://doi.acm.org/10.1145/3139958.3140022>.

MASRI, A.; ZEITOUNI, K.; KEDAD, Z.; LEROY, B. An automatic matcher and linker for transportation datasets. *ISPRS International Journal of Geo-Information*, Multidisciplinary Digital Publishing Institute, v. 6, n. 1, p. 29, 2017.

MCDONALD, G.; MACDONALD, C.; OUNIS, I. Enhancing sensitivity classification with semantic features using word embeddings. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2017. p. 450–463.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: *Proc. of NIPS 2013*. [S.l.: s.n.], 2013.

MIKUSZ, M.; BATES, O.; CLINCH, S.; DAVIES, N.; FRIDAY, A.; NOULAS, A. Understanding mobile user interactions with the IoT. In: *Proc. of MobiSys 2016*. [S.l.: s.n.], 2016. p. 140.

MINETT, P.; PEARCE, J. *Estimating the Energy Consumption Impact of Casual Carpooling*. 2011. 126–139 p.

MO, Y.; LI, B.; WANG, B.; YANG, L. T.; XU, M. Event recommendation in social networks based on reverse random walk and participant scale control. *FGCS*, Elsevier, v. 79, n. P1, p. 383–395, 2018.

MOHLER, B. J.; THOMPSON, W. B.; CREEM-REGEHR, S. H.; PICK, H. L.; WARREN, W. H. Visual flow influences gait transition speed and preferred walking speed. *Experimental Brain Research*, v. 181, n. 2, p. 221–228, 2007. ISSN 1432-1106. Disponível em: <http://dx.doi.org/10.1007/s00221-007-0917-0>.

ONAN, A. A machine learning based approach to identify geo-location of twitter users. In: *Proc. of ACM ICC 2017*. [S.l.: s.n.], 2017.

PARENT, C.; SPACCAPIETRA, S.; RENSO, C.; ANDRIENKO, G.; ANDRIENKO, N.; BOGORNY, V.; DAMIANI, M. L.; GKOULALAS-DIVANIS, A.; MACEDO, J. A.; PELEKIS, N.; THEODORIDIS, Y.; YAN, Z. Semantic trajectories modeling and analysis. *ACM Computing Surveys*, v. 45, n. 4, 2013.

PELEKIS, N.; THEODORIDIS, Y. *Mobility Data Management and Exploration*. Springer, 2014. ISBN 978-1-4939-0391-7. Disponível em: <https://doi.org/10.1007/978-1-4939-0392-4>.

PFRIEMER, H. The digital economy and the promise of a new mobility. In: _____. *Smart Mobility – Connecting Everyone: Trends, Concepts and Best Practices*. Wiesbaden: Springer Fachmedien Wiesbaden, 2017. p. 69–73. ISBN 978-3-658-15622-0. Disponível em: <https://doi.org/10.1007/978-3-658-15622-0_4>.

QIN, Z.; RISHABH, I.; CARNAHAN, J. A scalable approach for periodical personalized recommendations. In: *Proc. of ACM RecSys 2016*. [S.l.: s.n.], 2016.

QUERCIA, D.; LATHIA, N.; CALABRESE, F.; LORENZO, G. D.; CROWCROFT, J. Recommending social events from mobile phone location data. In: *Proc. of ICDM 2010*. [S.l.: s.n.], 2010.

RODE, P.; HOFFMANN, C.; KANDT, J.; GRAFF, A.; SMITH, D. Towards new urban mobility: the case of london and berlin. The London School of Economics and Political Science, LSE Cities, 2015.

SEOL, B. *Finding Similar Venues in Foursquare*. [S.l.]: Foursquare-Direct, 2015. Https://medium.com/foursquare-direct/finding-similar-venues-in-foursquare-cf535d9028ee.

SINNOTT, R. O.; CHEN, W. Estimating crowd sizes through social media. In: *2016 IEEE International Conference on Pervasive Computing and Communication Workshops, PerCom Workshops 2016, Sydney, Australia, March 14-18, 2016*. [s.n.], 2016. p. 1–6. Disponível em: <https://doi.org/10.1109/PERCOMW.2016.7457123>.

SINNOTT, R. O.; CHEN, W. Estimating crowd sizes through social media. In: *Proc. of IEEE PerCom Workshops 2016*. [S.l.: s.n.], 2016.

SINNOTT, R. O.; WANG, W. Estimating micro-populations through social media analytics. *Social Network Analysis and Mining*, Springer, v. 7, n. 1, p. 13, 2017.

SLOAN, L.; MORGAN, J.; HOUSLEY, W.; WILLIAMS, M.; EDWARDS, A.; BURNAP, P.; RANA, O. Knowing the tweeters: Deriving sociologically relevant demographics from twitter. *Sociological Research Online*, v. 18, n. 3, p. 1–11, 2013.

SOKOLOVA, M.; HUANG, K.; MATWIN, S.; RAMISCH, J.; SAZONOVA, V.; BLACK, R.; ORWA, C.; OCHIENG, S.; SAMBULI, N. Topic modelling and event identification from twitter textual data. *arXiv preprint arXiv:1608.02519*, 2016.

SPICKERMANN, A.; GRIENITZ, V.; HEIKO, A. Heading towards a multimodal city of the future?: Multi-stakeholder scenarios for urban mobility. *Technological Forecasting and Social Change*, Elsevier, v. 89, p. 201–221, 2014.

STEG, L.; VLEK, C. Encouraging pro-environmental behaviour: An integrative review and research agenda. *Journal of Environmental Psychology*, v. 29, p. 309–317, 2009.

STIGLIC, M.; AGATZ, N.; SAVELSBERGH, M.; GRADISAR, M. The benefits of meeting points in ride-sharing systems. *Transportation Research Part B: Methodological*, v. 82, n. Supplement C, p. 36 – 53, 2015. ISSN 0191-2615. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0191261515002088>.

STIGLIC, M.; AGATZ, N.; SAVELSBERGH, M.; GRADISAR, M. Making dynamic ride-sharing work: The impact of driver and rider flexibility. *Transportation Research Part E: Logistics and Transportation Review*, v. 91, n. Supplement C, p. 190 – 207, 2016. ISSN 1366-5545. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1366554515303033>.

TEODOROVIĆ, D.; DELL'ORCO, M. Mitigating traffic congestion: solving the ride-matching problem by bee colony optimization. *Transportation Planning and Technology*, Taylor & Francis, v. 31, n. 2, p. 135–152, 2008.

TOMTOM. *TomTom traffic index*. 2017. <https://www.tomtom.com/en_gb/trafficindex/>. Last access April 18, 2017.

TOTH, P.; VIGO, D. *The vehicle routing problem*. [S.l.]: SIAM, 2002.

TOWNSEND, A. M. *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*. [S.l.]: W.W. Norton and Company, 2014.

TRASARTI, R.; GIANNOTTI, F.; NANNI, M. Mining Mobility User Profiles for CarPooling. In: *KDD 2011*. [S.l.: s.n.], 2011. p. 1190–1198. ISBN 9781450308137.

URRY, J. *Mobilities: new perspectives on transport and society*. [S.l.]: Routledge, 2016.

WANG, S.; WANG, Z.; LI, C.; ZHAO, K.; CHEN, H. Learn to recommend local event using heterogeneous social networks. In: *Proc. of APWeb 2016*. [S.l.: s.n.], 2016.

WANG, X.; DESSOUKY, M.; ORDONEZ, F. A pickup and delivery problem for ridesharing considering congestion. *Transportation Letters*, Taylor & Francis, v. 8, n. 5, p. 259–269, 2016.

WANG, Y.; KUTADINATA, R. J.; WINTER, S. Activity-based ridesharing: increasing flexibility by time geography. In: *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS 2016, Burlingame, California, USA, October 31 - November 3, 2016*. [s.n.], 2016. p. 1:1–1:10. Disponível em: <http://doi.acm.org/10.1145/2996913.2997002>.

WEGENER, M. The future of mobility in cities: Challenges for urban modelling. *Transport Policy*, Elsevier, v. 29, p. 275–282, 2013.

WEIKL, S.; BOGENBERGER, K. Relocation strategies and algorithms for free-floating car sharing systems. *IEEE Intelligent Transportation Systems Magazine*, v. 5, n. 4, p. 100–111, winter 2013. ISSN 1939-1390.

WU, F.; LI, Z. Where did you go: Personalized annotation of mobility records. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2016. p. 589–598.

YANG, D.; ZHANG, D.; ZHENG, V. W.; YU, Z. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, v. 45, n. 1, p. 129–142, Jan 2015. ISSN 2168-2216.

ZHANG, S.; LV, Q. Event organization 101: Understanding latent factors of event popularity. In: *Proc. of ICWSM 2017*. [S.l.: s.n.], 2017.

ZHANG, S.; LV, Q. Hybrid egu-based group event participation prediction in event-based social networks. *Knowledge-Based Systems*, Elsevier, v. 143, p. 19 – 29, 2018.

ZHANG, X.; ZHAO, J.; CAO, G. Who will attend? – predicting event attendance in event-based social network. In: *Proc. of IEEE MDM 2015*. [S.l.: s.n.], 2015.

Appendix

Table 17 – Ride requests Q and Points of Interest (POIs) distributions by place category for the NYC dataset.

Category	Q	**POIs**	**Category**	Q	**POIs**
Afghan Restaurant	3	3	Internet Cafe	4	3
African Restaurant	18	14	Italian Restaurant	1,315	595
American Restaurant	2,681	716	Japanese Restaurant	239	174
Animal Shelter	21	12	Jewelry Store	91	60
Antique Shop	25	11	Korean Restaurant	233	113
Aquarium	11	2	Latin American Rest.	206	106
Arcade	85	47	Laundry Service	450	256
Arepa Restaurant	13	6	Library	286	101
Argentinian Rest.	11	10	Light Rail	350	64
Art Gallery	352	257	Mac & Cheese Joint	14	7
Art Museum	274	34	Malaysian Restaurant	18	14
Arts & Crafts Store	114	73	Mall	898	86
Arts & Entertainment	20	19	Market	92	3
Asian Restaurant	305	182	Medical Center	2,310	867
Athletic & Sport	487	226	Mediterranean Rest.	92	67
Australian Rest.	49	17	Mexican Restaurant	1,468	374
Automotive Shop	464	247	Middle Eastern Rest.	109	71
BBQ Joint	346	105	Military Base	9	8
Bagel Shop	595	209	Miscellaneous Shop	667	309
Bakery	899	372	Mobile Phone Shop	13	14
Bank	1,832	671	Molecular Gastro Rest.	7	3
Bar	11,242	2488	Moroccan Restaurant	8	5
Beach	341	76	Mosque	20	14
Beer Garden	318	56	Motorcycle Shop	1	1
Bike Rental/Share	12	5	Movie Theater	1401	139
Bike Shop	134	70	Moving Target	435	149
Board Shop	16	11	Museum	131	45
Bookstore	534	82	Music Store	54	24
Bowling Alley	133	41	Music Venue	1,027	192

Category	Q	**POIs**	**Category**	Q	**POIs**
Brazilian Restaurant	63	28	Nail Salon	65	48
Breakfast Spot	368	115	Newsstand	7	9
Brewery	88	18	Nightlife Spot	14	10
Bridal Shop	27	14	Other Great Outdoors	2,734	687
Burger Joint	1,047	295	Other Nightlife	78	58
Burrito Place	163	27	Outdoors & Recr.	173	50
Café	918	347	Office Supplies Store	261	88
Cajun/Creole Rest.	40	10	Park	3,015	536
Camera Store	31	8	Parking	281	97
Campground	159	13	Performing Arts Venue	464	126
Candy Store	118	49	Peruvian Restaurant	12	11
Car Dealership	29	13	Pet Service	2	3
Car Wash	42	15	Pet Store	260	105
Caribbean Restaurant	202	112	Photography Lab	2	2
Casino	71	8	Pizza Place	1394	843
Castle	5	1	Planetarium	11	4
Cemetery	178	40	Playground	313	184
Chinese Restaurant	932	451	Plaza	1,074	135
Church	730	317	Pool	77	41
Clothing Store	2,123	996	Pool Hall	55	26
Coffee Shop	4,106	853	Portuguese Restaurant	6	6
Comedy Club	89	40	Post Office	508	189
Concert Hall	109	44	Prof. & Other Places	153	53
Convenience Store	484	147	Public Art	24	11
Convention Center	227	64	Racetrack	69	20
Cosmetics Shop	412	262	Ramen/Noodle House	208	58
Cuban Restaurant	191	66	Record Shop	94	71
Cupcake Shop	185	50	Recycling Facility	5	3
Deli / Bodega	1,621	704	Rental Car Location	41	28
Department Store	1,386	185	Rest Area	163	26

Category	Q	**POIs**	**Category**	Q	**POIs**
Design Studio	76	44	Restaurant	324	180
Dessert Shop	312	142	River	64	13
Dim Sum Restaurant	49	10	Salad Place	112	32
Diner	1000	321	Salon / Barbershop	896	515
Distillery	7	5	Sandwich Place	1,030	411
Donut Shop	824	203	Scandinavian Rest.	10	7
Drugstore / Pharmacy	1,753	605	Scenic Lookout	543	139
Dumpling Restaurant	23	14	Science Museum	81	18
Eastern Euro Rest.	57	29	Sculpture Garden	105	44
Electronics Store	976	244	Seafood Restaurant	379	155
Embassy / Consulate	24	20	Shop & Service	28	24
Ethiopian Restaurant	9	11	Shrine	3	3
Event Space	314	201	Ski Area	29	28
Factory	56	21	Smoke Shop	68	29
Fair	17	7	Snack Place	48	29
Falafel Restaurant	64	31	Soup Place	103	32
Fast Food Restaurant	1215	433	South American Rest.	53	34
Ferry	532	91	Sout./Soul Food Rest.	108	49
Filipino Restaurant	22	9	Spa / Massage	411	221
Financ. or Legal Ser.	28	27	Spanish Restaurant	139	103
Fish & Chips Shop	5	5	Spiritual Center	31	25
Flea Market	214	37	Sporting Goods Shop	282	105
Flower Shop	44	35	Stadium	873	76
Food	105	42	Steakhouse	220	108
Food & Drink Shop	4693	1209	Storage Facility	25	6
Food Truck	524	234	Student Center	237	42
Fraternity House	29	29	Sushi Restaurant	448	289
French Restaurant	379	181	Swiss Restaurant	8	1
Fried Chicken Joint	241	129	Synagogue	116	38
Funeral Home	57	12	Taco Place	167	67
Furniture Store	365	124	Tanning Salon	40	22

Category	Q	**POIs**	**Category**	Q	**POIs**
Gaming Cafe	9	6	Tapas Restaurant	93	53
Garden	94	51	Tattoo Parlor	52	42
Garden Center	23	5	Taxi	124	74
Gas Station/Garage	941	390	Tea Room	176	64
Gastropub	141	48	Temple	19	11
General Entert.	882	316	Thai Restaurant	294	177
General Travel	1195	208	Theater	591	267
German Restaurant	72	39	Thrift / Vintage Store	127	66
Gift Shop	113	79	Toy / Game Store	197	49
Gluten-free Rest.	3	4	Travel & Transport	51	8
Greek Restaurant	82	57	Travel Lounge	14	8
Harbor / Marina	164	72	Turkish Restaurant	8	8
Hardware Store	377	127	Vegetar./Vegan Rest.	251	104
Historic Site	107	46	Video Game Store	88	57
History Museum	82	37	Video Store	27	21
Hobby Shop	37	15	Vietnamese Rest.	99	54
Hot Dog Joint	178	45	Winery	15	16
Housing Development	85	50	Wings Joint	98	38
Ice Cream Shop	467	221	Zoo	97	32
Indian Restaurant	226	156			

Table 18 – Ride requests Q and Points of Interest (POIs) distributions by place category for the TKY dataset.

Category	Q	**POIs**	**Category**	Q	**POIs**
Afghan Restaurant	1	1	Internet Cafe	164	64
African Restaurant	4	5	Italian Restaurant	2,345	1,164
American Restaurant	354	156	Japanese Restaurant	9,365	5,555
Animal Shelter	4	5	Jewelry Store	85	53
Antique Shop	37	20	Korean Restaurant	548	406
Aquarium	92	6	Latin American Rest.	5	3
Arcade	3,662	271	Laundry Service	110	84
Arepa Restaurant	1	1	Library	709	178
Art Gallery	731	275	Light Rail	2318	83
Art Museum	378	44	Mac & Cheese Joint	2	1
Arts & Crafts Store	333	116	Malaysian Restaurant	10	5
Arts & Entertainment	15	8	Mall	6,185	473
Asian Restaurant	682	383	Market	2	3
Athletic & Sport	457	214	Medical Center	2136	1,115
Australian Rest.	23	9	Mediterranean Rest.	24	16
Automotive Shop	402	180	Mexican Restaurant	92	43
BBQ Joint	1,234	738	Middle Eastern Rest.	45	11
Bagel Shop	31	26	Military Base	130	38
Bakery	857	458	Miscellaneous Shop	3,152	853
Bank	902	537	Mobile Phone Shop	980	329
Bar	8,051	4048	Moroccan Restaurant	3	3
Beach	40	11	Mosque	6	5
Beer Garden	252	81	Motorcycle Shop	16	8
Bike Rental/Share	13	8	Movie Theater	1802	134
Bike Shop	518	171	Moving Target	122	36
Board Shop	26	8	Museum	239	57
Bookstore	4,797	632	Music Store	210	84
Bowling Alley	535	47	Music Venue	1,598	340
Brazilian Restaurant	48	8	Nail Salon	35	17

Category	Q	**POIs**	**Category**	Q	**POIs**
Breakfast Spot	32	19	Newsstand	5	7
Brewery	101	33	Nightlife Spot	18	19
Bridal Shop	13	13	Other Great Outdoors	2139	609
Burger Joint	888	221	Other Nightlife	299	164
Burrito Place	6	3	Outdoors & Recr.	27	20
Café	5,967	2190	Office Supplies Store	213	69
Cajun/Creole Rest.	12	6	Park	4,026	942
Camera Store	261	44	Performing Arts Venue	73	35
Campground	22	11	Peruvian Restaurant	5	4
Candy Store	264	135	Pet Service	2	2
Car Dealership	94	54	Pet Store	129	60
Car Wash	17	10	Photography Lab	18	15
Caribbean Restaurant	15	6	Pizza Place	313	168
Casino	138	62	Planetarium	24	5
Castle	11	2	Playground	638	286
Cemetery	159	47	Plaza	1,030	98
Chinese Restaurant	2,804	1,607	Pool	74	38
Church	131	50	Pool Hall	73	36
Clothing Store	2,210	1,006	Portuguese Restaurant	7	4
Coffee Shop	4,756	1,167	Post Office	884	360
Comedy Club	95	22	Prof. & Other Places	54	28
Concert Hall	1,307	110	Public Art	43	11
Convenience Store	7,360	3,177	Racetrack	214	18
Convention Center	1,958	144	Ramen / Noodle House	10,618	3,609
Cosmetics Shop	132	95	Record Shop	1,159	114
Cuban Restaurant	4	1	Recycling Facility	76	16
Cupcake Shop	30	31	Rental Car Location	47	34
Deli / Bodega	623	327	Rest Area	164	36
Department Store	3,023	185	Restaurant	2,169	1,020
Design Studio	60	38	River	295	36

Category	Q	**POIs**	**Category**	Q	**POIs**
Dessert Shop	690	424	Salad Place	16	11
Dim Sum Restaurant	50	29	Salon / Barbershop	626	414
Diner	1,350	561	Sandwich Place	340	94
Distillery	6	6	Scandinavian Rest.	12	3
Donut Shop	373	134	Scenic Lookout	285	62
Drugstore / Pharmacy	1,236	693	Science Museum	142	24
Dumpling Restaurant	617	113	Sculpture Garden	265	54
Eastern Euro Rest.	22	15	Seafood Restaurant	345	179
Electronics Store	8,397	353	Shop & Service	194	89
Embassy / Consulate	110	56	Shrine	1,535	341
Ethiopian Restaurant	3	2	Ski Area	1	1
Event Space	1,658	349	Smoke Shop	533	135
Factory	147	59	Snack Place	112	74
Fair	70	26	Sorority House	3	4
Fast Food Rest.	3,698	1078	Soup Place	121	40
Ferry	198	47	South American Rest.	6	1
Financ. or Legal Serv.	12	7	Spa / Massage	672	321
Fish & Chips Shop	6	3	Spanish Restaurant	157	105
Flea Market	6	4	Spiritual Center	271	59
Flower Shop	59	53	Sporting Goods Shop	641	227
Food	194	147	Stadium	1,322	59
Food & Drink Shop	6,766	2136	Steakhouse	626	334
Food Truck	87	33	Storage Facility	5	7
Fraternity House	13	3	Student Center	143	25
French Restaurant	392	255	Sushi Restaurant	1,274	592
Fried Chicken Joint	301	148	Swiss Restaurant	1	2
Funeral Home	33	17	Synagogue	5	2
Furniture Store	894	247	Taco Place	7	6
Gaming Cafe	77	24	Tanning Salon	18	4
Garden	302	75	Tapas Restaurant	35	23
Garden Center	3	3	Taxi	52	17

Category	Q	**POIs**	**Category**	Q	**POIs**
Gas Station / Garage	542	224	Tea Room	151	93
Gastropub	83	33	Temple	1,171	316
General Entert.	831	193	Thai Restaurant	398	162
General Travel	378	104	Theater	519	108
German Restaurant	101	42	Thrift / Vintage Store	79	36
Gift Shop	466	150	Toy / Game Store	1,064	125
Gluten-free Rest.	1	1	Travel & Transport	104	27
Harbor / Marina	42	17	Travel Lounge	47	11
Hardware Store	394	67	Turkish Restaurant	13	10
Historic Site	919	229	Vegetar./Vegan Rest.	49	36
History Museum	163	53	Video Game Store	274	36
Hobby Shop	3,679	295	Video Store	1,071	197
Hot Dog Joint	25	8	Vietnamese Rest.	77	39
Housing Development	5	3	Winery	29	27
Ice Cream Shop	161	107	Wings Joint	149	56
Indian Restaurant	2,130	705	Zoo	93	25

Table 19 – Complement to Table 9 for Creamfields : accuracy achieved by all classifiers trained with BoW, w2v and both BoW+w2v features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Dataset	**Creamfields**						
Task	Model	Accuracy		Precision	Recall	F1-score	AuC
Before	$GBDT^{bow}$	0.874		0.846	0.912	0.878	0.873
	$GBDT^{w2v}_{sum}$	0.874	(0.0%)	0.869	0.874	0.871	0.874
	$GBDT^{bow}_{mean}$	0.872	(0.0%)	0.865	0.869	0.867	0.872
	LR^{bow}	0.868		0.870	0.870	0.868	0.887
	LR^{w2v}_{mix}	0.885	(+1.7%)	0.895	0.905	0.900	0.902
	LR^{both}_{max}	**0.913***	(+4.5%)	**0.927**	0.905	**0.916**	**0.919**
	NB^{bow}	0.587		0.540	0.977	0.696	0.600
	NB^{w2v}_{max}	0.585	(0.0%)	0.538	**0.982**	0.695	0.598
	NB^{both}_{mean}	0.583	(0.0%)	0.537	0.977	0.693	0.596
	RF^{bow}	0.826		0.760	0.941	0.840	0.830
	RF^{w2v}_{mix}	0.865*	(+3.9%)	0.842	0.897	0.867	0.866
	RF^{both}_{mix}	0.859*	(+3.3%)	0.834	0.892	0.861	0.860
	SVM^{bow}	0.607		0.591	0.599	0.593	0.606
	SVM^{w2v}_{sum}	0.637*	(+3.0%)	0.613	0.676	0.642	0.638
	SVM^{both}_{mix}	0.654*	(+4.8%)	0.628	0.698	0.661	0.656
During	$GBDT^{bow}$	0.817		0.830	0.616	0.708	0.790
	$GBDT^{w2v}_{max}$	0.789	(0.0%)	0.791	0.661	0.714	0.768
	$GBDT^{both}_{max}$	0.796	(0.0%)	0.796	0.667	0.720	0.773
	LR^{bow}	0.741		0.766	0.538	0.602	0.690
	LR^{w2v}_{sum}	0.804*	(+5.9%)	0.803	0.678	0.730	0.782
	LR^{both}_{mix}	**0.815***	(+7.0%)	0.811	**0.706**	**0.751**	**0.796**
	NB^{bow}	0.628		0.619	0.117	0.193	0.537
	NB^{w2v}_{mix}	0.637	(+0.9%)	0.816	0.117	0.195	0.544
	NB^{bow}_{mix}	0.637	(+0.9%)	0.816	0.117	0.195	0.544
	RF^{bow}	0.620		0.600	0.028	0.053	0.514
	RF^{w2v}_{mean}	0.780*	(+16.1%)	**0.855**	0.539	0.656	0.737
	RF^{both}_{mean}	0.752*	(+13.3%)	**0.885**	0.428	0.571	0.694
	SVM^{bow}	0.641		0.584	0.300	0.394	0.580
	SVM^{w2v}_{mix}	0.641	(0.0%)	0.584	0.289	0.383	0.578
	SVM^{both}_{mean}	0.643	(+0.2%)	0.591	0.300	0.396	0.582
After	$GBDT^{bow}$	0.780		0.792	0.948	0.864	0.640
	$GBDT^{w2v}_{max}$	0.830*	(+5.0%)	0.831	0.953	0.887	0.753
	$GBDT^{both}_{max}$	0.833*	(+5.3%)	0.836	0.947	0.888	0.760
	LR^{bow}	0.813		0.810	0.958	0.880	0.762
	LR^{w2v}_{sum}	0.824*	(+1.1%)	0.831	0.937	0.880	0.748
	LR^{both}_{sum}	**0.839***	(+2.6%)	**0.847**	0.937	**0.890**	**0.777**
	NB^{bow}	0.702		0.711	0.962	0.818	0.538
	NB^{w2v}_{mean}	0.704	(+0.2%)	0.710	0.969	0.820	0.537
	NB^{both}_{mean}	0.707	(+0.5%)	0.712	0.969	0.821	0.541
	RF^{bow}	0.713		0.708	**1.000**	0.829	0.532
	RF^{w2v}_{max}	0.780*	(+6.7%)	0.763	0.994	0.863	0.646
	RF^{both}_{mix}	0.770*	(+5.7%)	0.753	0.997	0.858	0.626
	SVM^{bow}	0.707		0.706	0.991	0.824	0.527
	SVM^{w2v}_{mix}	0.713	(+0.6%)	0.709	0.997	0.828	0.533
	SVM^{bow}_{mix}	0.713	(+0.6%)	0.708	**1.000**	0.829	0.532

Table 20 – Complement to Table 9 for VFestival: accuracy of all classifiers trained with BoW, w2v and both BoW+w2v features. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Dataset	**VFestival**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	$GBDT^{bow}$	0.809		0.802	0.768	0.784	0.808
	$GBDT^{w2v}_{max}$	0.818	(+0.9%)	0.832	0.776	0.802	0.816
	$GBDT^{both}_{max}$	**0.824**	(+1.5%)	0.804	0.835	**0.819**	**0.824**
	LR^{bow}	0.761		0.744	0.762	0.748	0.764
	LR^{w2v}_{mix}	0.778	(+1.3%)	0.793	0.826	0.807	0.814
	LR^{both}_{sum}	0.813*	(+4.8%)	0.792	0.826	0.807	0.813
	NB^{bow}	0.535		0.506	0.977	0.667	0.555
	NB^{w2v}_{mean}	0.535	(0.0%)	0.506	**0.982**	0.668	0.555
	NB^{both}_{sum}	0.535	(0.0%)	0.506	**0.982**	0.668	0.555
	RF^{bow}	0.778		**0.860**	0.648	0.735	0.772
	RF^{w2v}_{max}	0.804*	(+2.6%)	0.796	0.794	0.794	0.804
	RF^{both}_{mix}	0.798	(+2.0%)	0.787	0.799	0.790	0.798
	SVM^{bow}	0.578		0.568	0.471	0.514	0.573
	SVM^{w2v}_{mix}	0.609*	(+3.0%)	0.610	0.493	0.545	0.603
	SVM^{both}_{sum}	0.602*	(+2.4%)	0.603	0.484	0.537	0.597
During	$GBDT^{bow}$	0.802		0.850	0.582	0.688	0.763
	$GBDT^{w2v}_{max}$	0.823*	(+2.1%)	0.867	0.633	**0.727**	0.785
	$GBDT^{both}_{max}$	**0.826***	(+2.4%)	0.893	0.622	**0.727**	**0.787**
	LR^{bow}	0.626		0.600	0.614	0.494	0.606
	LR^{w2v}_{mix}	0.772*	(+14.6%)	0.855	0.500	0.626	0.722
	LR^{both}_{mix}	0.787*	(+16.1%)	0.887	0.505	0.639	0.732
	NB^{bow}	0.530		0.429	0.737	0.525	0.571
	NB^{w2v}_{mix}	0.433	(0.0%)	0.388	**0.895**	0.540	0.526
	NB^{both}_{mix}	0.446	(0.0%)	0.390	0.866	0.537	0.530
	RF^{bow}	0.680		**1.000**	0.145	0.248	0.573
	RF^{w2v}_{sum}	0.796*	(+11.5%)	0.907	0.512	0.651	0.740
	RF^{both}_{max}	0.754*	(+7.4%)	0.845	0.442	0.576	0.695
	SVM^{bow}	0.670		0.800	0.157	0.257	0.566
	SVM^{w2v}_{sum}	0.676	(0.7%)	0.790	0.175	0.281	0.575
	SVM^{both}_{mean}	0.670	(0.00%)	0.800	0.157	0.257	0.566
After	$GBDT^{bow}$	0.815		0.824	0.902	0.862	0.767
	$GBDT^{w2v}_{mix}$	**0.861***	(+4.6%)	0.862	0.945	**0.901**	**0.817**
	$GBDT^{both}_{mix}$	0.854*	(+3.9%)	0.848	0.948	0.894	0.799
	LR^{bow}	0.809		0.812	0.932	0.868	0.808
	LR^{w2v}_{mix}	0.850*	(+4.1%)	0.858	0.932	0.893	0.807
	LR^{both}_{sum}	0.858*	(+4.9%)	**0.877**	0.919	0.897	0.827
	NB^{bow}	0.696		0.709	0.929	0.804	0.574
	NB^{w2v}_{mix}	0.717*	(+2.1%)	0.717	0.958	0.820	0.592
	NB^{both}_{mix}	0.717*	(+2.1%)	0.717	0.958	0.820	0.592
	RF^{bow}	0.689		0.684	**1.000**	0.812	0.527
	RF^{w2v}_{sum}	0.789*	(+10.0%)	0.782	0.951	0.858	0.704
	RF^{both}_{mix}	0.774*	(+8.5%)	0.763	0.964	0.851	0.674
	SVM^{bow}	0.707		0.699	0.994	0.820	0.556
	SVM^{w2v}_{mean}	0.709	(+0.2%)	0.705	0.977	0.819	0.568
	SVM^{both}_{mean}	0.709	(+0.2%)	0.699	0.997	0.822	0.558

Table 21 – Complement to Table 11 on generalization ability of the various classifiers: models trained on Creamsfields are tested on VFestival. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**Creamfields/VFestival**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	$GBDT^{bow}$	0.780	(0.0%)	0.757	0.795	0.775	**0.862**
	LR^{both}_{mix}	**0.796**	(+1.3%)	**0.783**	0.790	**0.786**	0.861
	NB^{bow}_{mean}	0.546	(+0.3%)	0.512	**0.945**	0.665	0.565
	RF^{bow}	0.778	(0.0%)	0.748	0.758	0.753	0.843
	SVM^{both}_{mix}	0.526	(+0.7%)	0.502	0.470	0.486	0.497
During	$GBDT^{w2v}_{max}$	**0.724***	(+1.3%)	**0.619**	**0.680**	**0.648**	**0.797**
	LR^{both}_{max}	0.702*	(+3.2%)	0.607	0.576	0.591	0.732
	NB^{both}_{mix}	0.524*	(+4.1%)	0.247	0.134	0.174	0.419
	RF^{both}_{max}	0.693*	(+7.8%)	0.604	0.523	0.561	0.679
	SVM^{w2v}_{mix}	0.643*	(+1.7%)	0.583	0.163	0.255	0.520
After	$GBDT^{w2v}_{sum}$	**0.789***	(+5.6%)	0.769	0.981	**0.862**	**0.845**
	LR^{both}_{mix}	0.787*	(+6.7%)	**0.773**	0.968	0.859	0.817
	NB^{both}_{mean}	0.698	(0.0%)	0.699	0.968	0.811	0.559
	RF^{w2v}_{sum}	0.735*	(+4.4%)	0.747	0.916	0.823	0.754
	SVM^{w2v}_{mix}	0.667*	(+1.5%)	0.671	**0.990**	0.800	0.549

Table 22 – Complement to Table 11 on generalization ability of the various classifiers: models trained on VFestival are tested on Creamsfields and vice versa. The * indicates statistical significant differences compared to the best classifiers using only BoW features (McNemar's test with 95% confidence interval).

Training/Test	**VFestival/Creamfields**						
Task	Model	Accuracy		Precision	Recall	F1	AuC
Before	$GBDT^{bow}$	0.824	(0.0%)	0.844	0.779	0.810	0.912
	LR^{w2v}_{mix}	**0.865***	(+1.3%)	**0.867**	0.851	**0.859**	**0.920**
	NB^{bow}	0.570	(0.0%)	0.529	**0.991**	0.690	0.586
	RF^{bow}	0.808	(0.0%)	0.876	0.667	0.757	0.886
	SVM^{both}_{mix}	0.546	(+0.4%)	0.541	0.387	0.451	0.486
During	$GBDT^{both}_{max}$	**0.743***	(+3.2%)	**0.810**	0.450	**0.579**	**0.796**
	LR^{both}_{sum}	0.741*	(+9.2%)	0.802	0.450	0.577	0.803
	NB^{both}_{sum}	0.370	(+0.3%)	0.377	**0.933**	0.537	0.468
	RF^{bow}	0.678	(0.0%)	0.686	0.328	0.444	0.677
	SVM^{w2v}_{sum}	0.593*	(+5.7%)	0.370	0.056	0.097	0.507
After	$GBDT^{both}_{mean}$	**0.807***	(+3.7%)	0.844	0.884	**0.864**	**0.863**
	LR^{bow}	0.787	(0.0%)	**0.862**	0.824	0.843	0.857
	NB^{bow}	0.709	(+0.2%)	0.717	0.959	0.820	0.550
	RF^{w2v}_{sum}	0.726*	(+3.7%)	0.818	0.777	0.797	0.757
	SVM^{bow}	0.689	(0.0%)	0.699	**0.969**	0.812	0.582

Table 23 – Complement to Table 12: robustness of the GBDT, LR and RF classifiers exploiting NFV features. Models trained on Creamsfields are tested on VFestival. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 21 (McNemar's test with 95% of confidence interval). Results of NB and SVM classifiers are not reported since they do not improve by using the NFV features.

Train/Test	**Creamfields/VFestival**						
Task	$Model_{aggv,nfv^{(top)}}$	Accuracy		Precision	Recall	F1	AuC
Before	$GBDT^{w2v}_{mix,mean^{(3)}}$	0.793	(+0.4%)	0.772	**0.804**	0.787	**0.867**
	$LR^{both}_{max,sum^{(3)}}$	**0.800**	(+0.4%)	**0.787**	0.795	**0.791**	0.861
	RF^{bow}	0.763	(0.00%)	0.752	0.749	0.751	0.831
During	$GBDT^{w2v}_{max,max^{(1)}}$	**0.746***	(+2.2%)	0.646	**0.709**	**0.676**	**0.817**
	$LR^{both}_{max,max^{(1)}}$	0.707	(+0.5%)	0.612	0.587	0.599	0.732
	$RF^{both}_{max,sum^{(1)}}$	0.713	(+2.0%)	**0.647**	0.512	0.571	0.733
After	$GBDT^{both}_{sum,sum^{(5)}}$	**0.811***	(+2.2%)	**0.792**	0.974	**0.874**	**0.872**
	$LR^{both}_{mix,max^{(1)}}$	**0.811**	(+2.4%)	0.789	**0.981**	0.874	0.866
	$RF^{both}_{sum,sum^{(1)}}$	0.759	(+2.4%)	0.783	0.887	0.832	0.794

Table 24 – Complement to Table 12: robustness of the GBDT, LR and RF classifiers exploiting NFV features. Models trained on VFestival are tested on Creamsfields. The * indicates statistically significant improvements with respect to the best accuracy figures reported in Table 22 (McNemar's test with 95% of confidence interval). Results of NB and SVM classifiers are not reported since they do not improve by using the NFV features.

Train/Test	**Creamfields/VFestival**						
Task	$\text{Model}_{\text{aggv,nfv}^{(\text{top})}}$	Accuracy		Precision	Recall	F1-Score	AuC
Before	$\text{GBDT}^{\text{w2v}}_{\text{sum,sum}^{(1)}}$	0.861	(+2.6%)	0.891	0.811	0.849	0.917
	$\text{LR}^{\text{both}}_{\text{max,sum}^{(3)}}$	**0.872***	(+0.7%)	0.860	**0.860**	**0.860**	0.915
	$\text{RF}^{\text{w2v}}_{\text{none,mix}^{(1)}}$	0.833	(+2.4%)	**0.919**	0.716	0.805	**0.921**
During	$\text{GBDT}^{\text{both}}_{\text{mean,mix}^{(1)}}$	**0.778***	(+3.5%)	**0.792**	0.550	**0.649**	**0.828**
	$\text{LR}^{\text{both}}_{\text{sum,max}^{(1)}}$	0.757*	(+1.6%)	0.758	**0.556**	0.641	0.797
	$\text{RF}^{\text{w2v}}_{\text{mix,max}^{(1)}}$	0.702	(+2.8%)	0.717	0.394	0.509	0.725
After	$\text{GBDT}^{\text{both}}_{\text{sum,mean}^{(5)}}$	**0.817***	(+1.0%)	0.840	0.903	**0.870**	0.839
	$\text{LR}^{\text{w2v}}_{\text{mean,sum}^{(3)}}$	0.811*	(+2.4%)	**0.847**	0.868	0.858	**0.855**
	$\text{RF}^{\text{both}}_{\text{sum,mean}^{(5)}}$	0.765*	(+7.6%)	0.771	**0.940**	0.847	0.788